UIUCDCS-R-76-820

Error Estimation and Iterative Improvement for the Numerical Solution of Operator Equations

by

Bengt Lindberg

July 1976

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · URBANA, ILLINOIS

UIUCDCS-R-76-820

Error Estimation and Iterative Improvement for the Numerical Solution of Operator Equations

by

Bengt Lindberg

July 1976

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801

Supported in part by the United States Office of Scientific Research under contract AFOSR-75-2854.

Acknowledgment

This work was supported by the Air Force Office of Scientific Research under grant AFOSR 75-2854. The author wishes to thank Professor Robert Skeel, who carefully read the original manuscript and made valuable suggestions for the improvement of it; Professor C. W. Gear who arranged my visit to the University of Illinois and suggested this line of research; Mr. John van Rosendale who programmed parts of the numerical experiments; Ms. Pamela Farr who typed the manuscript.

Abstract

A method for estimation of the global discretization error of solutions of operator equations is presented. Further an algorithm for iterative improvement of the approximate solution of such problems is given. The theoretical foundation for the algorithms are given as a number of theorems. Several classes of operator equations are examined and numerical results for both the error estimation algorithm and the algorithm for iterative improvement are given for some classes of ordinary and partial differential equations and integral equations.

Table of Contents

1. Introduction

A simple technique for estimating the discretization error, or improving the accuracy of the numerical solution of a functional equation

$$F(y) = 0$$

is discussed. The error estimate is obtained as the difference between the solution of the discretized problem

$$\phi_h(\eta) = 0$$

and a slightly perturbed discrete problem

$$\phi_h(\eta^E) + \phi_h^E(\eta) = 0 \ ;$$

the improved solutions are obtained from a sequence of perturbed problems.

The perturbation operator ϕ_h^E is a discrete approximation to the operator F that is of higher order of accuracy than ϕ_h (i.e., if ϕ_h is consistent of order p with F, then ϕ_h^E is consistent of order $q > p$ with F, see section 2 for more precise statements). We can view $-\phi_h^E(\eta)$ as an approximation to the local discretization error of the operator ϕ_h.

The operator ϕ_h^E corresponds to a more accurate discretization method than ϕ_h. The basic requirement on ϕ_h^E is the accuracy; the operator does not even need to be stable as it is applied only to the known quantity η to produce a constant to be added to ϕ_h.

The perturbations needed to iteratively improve the solution are obtained from operators that are increasingly accurate approximations to F, and the negative of the perturbation can again be viewed as increasingly accurate approximations to the local discretization error of the operator ϕ_h. If the order of accuracy of the operator ϕ_h is p we can gain p orders of accuracy per iteration if the perturbations are chosen judiciously.

One simple and direct way of constructing the perturbations is the following (other ways will be examined in section 4):

For the error estimation algorithm compute the residual (on a certain set of discrete points), which is obtained when the numerical solution η of the discrete problem $\phi_h(\eta) = 0$ (a solution defined only on a set of discrete points) is substituted for the unknown function y in the functional equation. Any functionals (e.g. derivatives) in $F(y) = 0$ that had to be approximated to get the discrete problem $\phi_h(\eta) = 0$ are approximated by linear combinations of the components of the vector η (solution of $\phi_h(\eta) = 0$) when the residual is computed.

To improve the solution the same technique is used, but now the residuals are computed as the sum of the old residuals and a new residual computed from the most recent approximation to the solution. Further increasingly accurate formulas are used to calculate necessary approximations to functionals that appear in F(y).

This technique is useful if there exists an expansion of the global discretization error (in the discretization parameter h) of the discretized problem $\phi_h(\eta) = 0$ to sufficiently high order and with smooth error terms. If sufficient smoothness conditions are met (e.g. in those cases where iterated deferred corrections can be applied) several iterative improvements can be made, for each iteration the order of accuracy of the new approximate solution is increased, in the same way as when iterated deferred corrections are used.

For the cases where one cannot perform iterative improvement due to the existence of significant non-smooth terms in the global error expansion one may still, in many cases, get realistic error estimates with the technique described.

Related techniques are difference correction, Fox [1947], Volkov [1957], and iterated deferred corrections according to Pereyra. Pereyra has worked extensively on boundary value problems for ordinary differential equations, Pereyra [1967b, 1973] and has produced very efficient computer codes, Lentini, Pereyra [1974, 1975a, 1975b] for such problems. He has also treated linear and mildly nonlinear elliptic boundary value problems, Pereyra [1967b], Pereyra [1970] and analyzed general nonlinear functional equations, Pereyra [1967a].

The basic computational tools for high order approximations to linear functionals, needed both in deferred corrections and iterative improvement can be found in Ballester, Pereyra [1966], Björck, Pereyra [1970], Galimberti, Pereyra [1970].

Whenever iterated deferred corrections can be used, iterative improvement can also easily be used. However to be able to perform iterated deferred corrections one must know and be able to approximate the terms of the local error expansion for the discretization operator ϕ_h, while for iterative improvement that is not necessary. In some cases, e.g. for boundary value problems $y'' = f(x, y)$; $y(a) = \alpha$, $y(b) = \beta$, these local error expansions are easily found (by Taylor expansions) and involve no terms that are difficult to approximate. In other cases, e.g. for problems where $F(y)$ is nonlinear in some derivative of y, the expansions are very cumbersome to find and even more cumbersome to approximate (see Pereyra [1970]), which makes iterated deferred corrections, although theoretically feasible, in practice impossible to perform.

Another related technique is iterated defect corrections according to Stetter [1974], Frank [1975]. The main difference between the approach

of the two later papers and our approach is that they define a smooth global approximation to the solution of the operator equation $F(y) = 0$ from the discrete solution η of $\phi_h(\eta) = 0$ and then compute the perturbation from that global approximation, while we use the discrete solution directly and need to be concerned only about the local properties of the perturbations.

In section 3 of Stetter [1974] the author also introduces a formalism, similar to ours, in order to discuss how the leading term of the local discretization error for the numerical solution of ordinary differential equations usually is estimated. He does, however, not use this formalism any further.

The starting point for our investigation was Stetter [1974] and the many papers of Pereyra (1967), ..., (1975) and our primary goal was to find cheap ways to estimate the errors in the numerical solution of partial differential equations.

Extensive discussions of the existence of asymptotic expansions of global truncation errors for the numerical solution of functional equations can be found in Stetter [1965], Pereyra [1967a], and Stetter [1973].

2. General theory

2.1 Preliminaries

The notation of this section is greatly influenced by Stetter [1965, 1973] and Pereyra [1967a].

The basis for the results discussed here is the existence of asymptotic error expansions for the global discretization error of the solution to the discretized problem. For a thorough discussion of this matter, see the references above.

We consider functional equations

$$F(y) = 0 \tag{2.1}$$

where $F: E \to E^0$ is a generally nonlinear operator from a Banach space E into a Banach space E^0. We will always assume that (2.1) has a unique solution $y \in E$.

For the purpose of numerical solution the problem (2.1) is discretized in the following sense:

We define families - depending on a real parameter $h \in H$, $H = \{h = \frac{c}{n}$, n integer in a subset of $\{\nu, \nu \geq n_0\}$ with $n_0 > 0$ fixed$\}$ - of Banach spaces E_h, E_h^0 and of bounded linear transformations Δ_h, Δ_h^0, which map E, E^0 into E_h, E_h^0, respectively.

Then we choose a family of (nonlinear) operators $\phi_h: E_h \to E_h^0$ such that for $z \in E$ and $h \in H$ or for $z = y$ and $h \in H$

$$\phi_h(\Delta_h z) = \Delta_h^0\{F(z) + \sum_{\nu=p}^{M} h^\nu \cdot f_\nu(z)\} + O(h^{M+1}), \tag{2.2}$$

where $f_\nu: E \to E^0$ does not depend upon h.

If for all $z \in E$ $||\phi_h(\Delta_h z) - \Delta_h^0 F(z)||_{E_h^0} = O(h^p)$ the operator ϕ_h is said to be consistent of order p with F.

The expression $\phi_h(\Delta_h y)$ formed with the solution y of (2.1) is often called the local discretization error of ϕ_h.

The original problem (2.1) is now replaced by the "algorithm"

$$\phi_h(\eta) = 0, \tag{2.3}$$

which is supposed to have a unique solution $\eta(h) \in E_h$ for $h \in H$. See the references above for a discussion of the unique solvability of $\phi_h(\eta) = 0$. The global discretization error of (2.3) is defined as

$$\varepsilon(h) = \eta(h) - \Delta_h y \in E_h$$

where y is again the solution of (2.1).

(2.3) is convergent of order p ($p \geq 1$) if

$$||\varepsilon(h)||_{E_h} \leqq C\, h^p \quad \text{for} \quad h \in H.$$

The global discretization error $\varepsilon(h)$ admits an asymptotic expansion to the order M ($M \geq p$) if there are $e_v \in E$, $v = p(1)M$, e_v independent of h, such that

$$||\varepsilon(h) - \Delta_h \sum_{v=p}^{M} h^v \cdot e_v||_{E_h} \leqq C_M\, h^{M+1} \quad \text{for} \quad h \in H. \tag{2.5}$$

We will call ϕ_h stable at $\Delta_h z$ if there exist constants S and $r > 0$ such that uniformly for all $h \in H$

$$||\xi^1 - \xi^2||_{E_h} \leqq S[||\phi_h(\xi^1) - \phi_h(\xi^2)||_{E_h^0} \tag{2.6a}$$

for all ξ^i, $i = 1, 2$ such that

$$||\phi_h(\xi^i) - \phi_h(\Delta_h z)||_{E_h^0} < r \tag{2.6b}$$

(cf. def. 1.1.10 in Stetter [1973]).

To estimate the global truncation error choose a family of (non-linear) operators ϕ_h^E: $E_h \to E_h^0$ such that for $z \in E$ and $h \in H$

$$\phi_h^E(\Delta_h z) = \Delta_h^0\{F(z)\} + O(h^q) \quad ; \quad p < q \leq 2p \tag{2.7}$$

When the solution η of problem (2.3) is known, solve for η^E from

$$\phi_h(\eta^E) + \phi_h^E(\eta) = 0 \tag{2.8}$$

Under suitable assumptions

$$\eta - \eta^E = \Delta_h[\sum_{v=p}^{q-1} h^v e_v] + O(h^q) \tag{2.9}$$

In this report we are mainly interested in estimating the global truncation error for given algorithms, however the idea behind (2.7), (2.8) can be extended to iteratively compute better approximations to the solution of (2.1).

To iteratively compute better approximations to the solution of (2.1) choose a family of (non-linear) operators $\phi_{h,i}$: $E_h \to E_h^0$, $i = 1, 2, \ldots$ such that for $z \in E$ and $h \in H$ or for $z = y$ and $h \in H$

$$\phi_{h,i}(\Delta_h z) = \Delta_h^0\{F(z) + \sum_{v=(i+1)p}^{M} h^v f_{v,i}(z)\} + O(h^{M+1}) \tag{2.10}$$

Put $\eta^0 = \eta$ and compute η^i, $i = 1, 2, \ldots$ recursively from

$$\phi_h(\eta^i) + \sum_{v=1}^{i} \phi_{h,v}(\eta^{v-1}) = 0 \quad i = 1, 2, \ldots \tag{2.11}$$

Under suitable assumptions

$$\eta^i - \Delta_h y = \Delta_h[\sum_{v=(i+1)p}^{M_i} h^v e_{v,i}] + O(h^{M_i+1}) \tag{2.12}$$

A simple example of the use of the operator formalism of this section is given in note 5 after theorem 2. Readers that are unfamiliar with the notation of this section are advised to study that example before they read the theorems.

Note 1 $$\phi_h^E(\Delta_h\, y) = \Delta_h^0\{F(y)\} + O(h^q)$$

so ϕ_h is consistent with F of order q. Further, under suitable assumptions (see the theorems below)

$$\phi_h(\Delta_h\, y) + \phi_h^E(\eta) = \Delta_h^0\{F(y)\} + O(h^q)$$

so $\phi_h + \phi_h^E(\eta)$ is consistent with F of order q. We can view $-\phi_h^E(\eta)$ as an approximation, accurate to order q - 1 in h, to the local discretization error of the operator ϕ_h.

Note 2 $$\phi_{h,i}(\Delta_h\, y) = \Delta_h^0\{F(y)\} + O(h^{(i+1)\cdot p})$$

so $\phi_{h,i}$ is consistent of order $(i+1)\cdot p$ with F. Further, under suitable assumptions, (see theorem 3 below)

$$\phi_h(\Delta_h\, y) + \sum_{v=1}^{i} \phi_{h,v}(\eta^{v-1}) = \Delta_h^0\{F(y)\} + O(h^{(i+1)\cdot p})$$

(This is not proved in the theorem, but can easily be proved with the technique used in the proof of the theorem.) Thus $\phi_h + \sum_{v=1}^{i} \phi_{h,v}(\eta^{v-1})$ is consistent of order $(i+1)\cdot p$ with F. We can thus view $-\sum_{v=1}^{i} \phi_{h,v}(\eta^{v-1})$ as an approximation, accurate to order $(i+1)\cdot p - 1$ in h, to the local discretization error of the operator ϕ_h.

Note 3 From the consistency of the operators above and the stability of ϕ_h (which implies the stability of $\phi_h + g_h$ for any constant $g_h \in E_h^0$) one can derive results on the convergence and the existence of asymptotic error expansions for the algorithms (2.8) and (2.11).

2.2 Basic theorems

Theorem 1

Let y be the unique solution of $F(y) = 0$ and let

a) the global discretization error $\eta - \Delta_h y$ of (2.3) have an asymptotic expansion

$$\eta - \Delta_h y = \Delta_h \{ \sum_{j=p}^{M_0} h^j e_j \} + \delta^0(h)$$

with $M_0 \geq p + k$ and $||\delta^0(h)|| = O(h^{M_0+1})$

b) the expansions (2.2) and (2.7) hold with $M \geq p$ and $p < q \leq 2p$

c) the operator ϕ_h be stable at $\Delta_h y$ in the sense of (2.6)

d) there exist constants L and b such that uniformly for all $h \in H$

$$||f_j(y^1) - f_j(y^2)||_{E^0} \leq L||y^1 - y^2||_E$$

for all $y^i \in E$, $i = 1, 2$ such that

$$||y^i - y||_E \leq b$$

$j = p, p+1, \ldots, N_1$ ($N_1 = \min (M, M_0 - k, q - 1)$)

e) there exist constants C, C_E and d such that uniformly for all $h \in H$

$$||\phi_h(\xi^1) - \phi_h(\xi^2)||_{E_h^0} \leq C \cdot ||\xi^1 - \xi^2||_{E_h} \cdot h^{-k}$$

$$||\phi_h^E(\xi^1) - \phi_h^E(\xi^2)||_{E_h^0} \leq C_E ||\xi^1 - \xi^2||_{E_h} \cdot h^{-k}$$

for any ξ^1, $\xi^2 \in E_h$ such that

$$||\xi^i - \Delta_h y||_{E_h} \leq d \quad , \quad i = 1, 2$$

Then the solution η^E of

$$\phi_h(\eta^E) + \phi_h^E(\eta) = 0$$

satisfies the inequality

$$\eta^E = \Delta_h y + O(h^{N_1+1})$$

where $N_1 = \min [M, M_0 - k, q - 1]$.

Proof

Note that

$$0 = \phi_h(\eta^E) + \phi_h^E(\eta) = \phi_h(\eta^E) - \phi_h(\Delta_h y) + \phi_h(\Delta_h y) + \phi_h^E(\eta)$$

i.e.

$$\phi_h(\eta^E) - \phi_h(\Delta_h y) = - \phi_h(\Delta_h y) - \phi_h^E(\eta)$$

We will show below that $\phi_h(\Delta_h y) + \phi_h^E(\eta) = O(h^{N_1+1})$ so for sufficiently small h we have

$$||\phi_h(\eta^E) - \phi_h(\Delta_h y)|| < r$$

Thus from the stability of ϕ_h at $\Delta_h y$ we have

$$||\eta^E - \Delta_h y|| \leq S||\phi_h(\eta^E) - \phi_h(\Delta_h y)|| = S||- \phi_h(\Delta_h y) - \phi_h^E(\eta)||$$

Here and in the sequel, whenever there is no danger of confusion, we omit the indices on the norms that refer to the actual Banach spaces.

Introduce the notation

$$z = \sum_{j=p}^{M_0} h^j e_j$$

and form

$$\begin{aligned}
\phi_h(\Delta_h y) + \phi_h^E(\eta) &= \phi_h(\Delta_h y) - \phi_h(\eta) + \phi_h^E(\eta) \\
&= \phi_h(\Delta_h y) - \phi_h(\Delta_h(y + z) + \delta^0) + \phi_h^E(\Delta_h(y + z) + \delta^0) \\
&= \phi_h(\Delta_h y) - \phi_h(\Delta_h(y + z)) + \phi_h^E(\Delta_h(y + z)) + O(h^{M_0+1-k}) \\
&= \Delta_h^0\{F(y) + \sum_{j=p}^{N_1} f_j(y)\, h^j - F(y + z) - \sum_{j=p}^{N_1} f_j(y + z)\, h^j \\
&\quad + F(y + z)\} + O(h^{N_1+1}) \\
&= \Delta_h^0\{\sum_{j=p}^{N_1} [f_j(y) - f_j(y + z)]\, h^j\} + O(h^{N_1+1}) \\
&= O(h^{N_1+1})
\end{aligned}$$

Hence

$$\eta^E - \Delta_h y = O(h^{N_1+1})$$

Corollary

Let the conditions of theorem 1 be satisfied with $M_0 \geq 2p + k - 1$, $q = 2p$, $M \geq 2p - 1$, then the solution η^E of

$$\phi_h(\eta^E) + \phi_h^E(\eta) = 0$$

satisfies the inequality

$$\eta^E = \Delta_h y + O(h^{2p}).$$

■

Theorem 2

Let y be the unique solution of $F(y) = 0$ and

a) conditions a), c) and e) of theorem 1 hold

b) F, ϕ_h and ϕ_h^E be twice continuously Frechet differentiable

c) the inequalities below hold

$$\phi_h(\Delta_h\, y) = \Delta_h^0\{F(y) + \sum_{j=p}^{M} f_j(y)\, h^j\} + O(h^{M+1})$$

$$\phi_h^E(\Delta_h\, y) = \Delta_h^0\{F(y)\} + O(h^q) \qquad p < q \leq 2p$$

$$\phi_h'(\Delta_h\, y) = \Delta_h^0\{F'(y)\} + O(h^{q-p})$$

$$\phi_h^{E'}(\Delta_h\, y) = \Delta_h^0\{F'(y)\} + O(h^{q-p})$$

with $M \geq p$. Further let $\phi_h^{(2)}(\Delta_h\, z)$ and $\phi_h^{E^{(2)}}(\Delta_h\, z)$ be uniformly bounded with respect to h for any $z \in E$. Then the solution η^E of

$$\phi_h(\eta^E) + \phi_h^E(\eta) = 0$$

satisfies the inequality

$$\eta^E = \Delta_h\, y + O(h^{N_1+1})$$

$$N_1 = \min(M, M_0 - k, q - 1)$$

Proof

Use the same notation as in the proof of theorem 1. This proof proceeds in the same way as that proof, except that

$$\phi_h(\Delta_h\, y) + \phi_h^E(\eta) = \phi_h(\Delta_h\, y) - \phi_h(\Delta_h(y + z)) + \phi_h^E(\Delta_h(y + z)) + O(h^{M_0+1-k})$$

$$= \phi_h(\Delta_h\, y) - \phi_h(\Delta_h\, y) - \phi_h'(\Delta_h\, y)\, \Delta_h\, z + \phi_h^E(\Delta_h\, y)$$

$$+ \phi_h^{E'}(\Delta_h\, y)\, \Delta_h\, z + O(h^{N_1+1})$$

$$= \Delta_h^0\{- F'(y)\, z - (\sum_{j=p}^{M} f_j'(y)\, h^j)\, z + F(y) + F'(y)\, z\}$$

$$+ O(h^{N_1+1})$$

$$= O(h^{N_1+1})$$

■

Note 1 These theorems are the basis for estimation of global discretization errors for algorithms where the error expansion contains only a few smooth terms. If $k \geq 1$, and $p \leq M_0 < p + k$ the error expansions contain at least one smooth term $h^p e_p$ but due to condition e) no estimate of the error can be obtained. Numerical experiments (see section 4.5) indicate that it might be possible to relax this condition. No theoretical results along these lines have yet been obtained.

Note 2 The condition on M in b) of theorem 1 is somewhat less restrictive than the condition on M_0 in a). However, to obtain the asymptotic expansion in a) one would need to have $M \geq p + k$ in the expansion (2.2). The lower limit on M for the expansion (2.7), though, may be of value in some cases.

Note 3 The constant k of condition e) is in general the order of the highest derivative present in the functional equation F(y).

Note 4 These theorems differ only in the differentiability assumptions on the operators involved and the set of elements from E for which the inequalities (2.2) and (2.7) must be valid.

Due to the fact that (2.2) and (2.7) only are needed for the exact solution y to $F(y) = 0$ in theorem 2 some very interesting types of perturbation operators ϕ_h^E are allowed in that theorem, while they are not allowed in theorem 1. In essence, all operators ϕ_h^E for which the local discretization error is of order greater than p can be used in theorem 2, while in theorem 1 the order of consistency of ϕ_h^E with F must exceed p. Several examples in section 4 will clarify this distinction.

Whether the more relaxed differentiability conditions of theorem 1 are of any practical importance I don't know. Future studies will hopefully provide some answers to this question.

Note 5 The value of k of condition e) in theorem 1 depends on the norm for the space E_h^0. For linear multistep methods for initial value problems for ordinary differential equations one can (at least for ϕ_h) get $k = 0$ by choosing Spijker's norm (see Stetter (1973), section 2.2.4, p. 81-84) for E_h^0. To get $k = 0$ also for ϕ_h^E the operator must be chosen judiciously, cf. also note 2 on p. 41 of section 4.1.

Note 6 As an exercise in the formalism of this section we analyze an algorithm for the two-point boundary value problem

$$y'' = f(x, y)$$

$$y(a) = \alpha \quad ; \quad y'(b) = \beta \quad ; \quad f \in C^{2s}([a, \bar{b}] \times R) \quad ; \quad \bar{b} = b + \varepsilon$$

The 2s-continuity of f is chosen for convenience of notation.

1. Operator equation

$$F: \; E \to E^0$$

with

$$E = C^{2s}[a, \bar{b}]$$

$$E^0 = \mathbb{R} \times C[a, \bar{b}] \times \mathbb{R}$$

and the following norms

$$||z||_E = \max_{a \leq x \leq \bar{b}} \sum_{v=0}^{2s} |z^{(v)}(x)|/v!$$

$$||g||_{E^0} = |g^1| + |g^2| + \max_{a \leq x \leq \bar{b}} |g(x)|$$

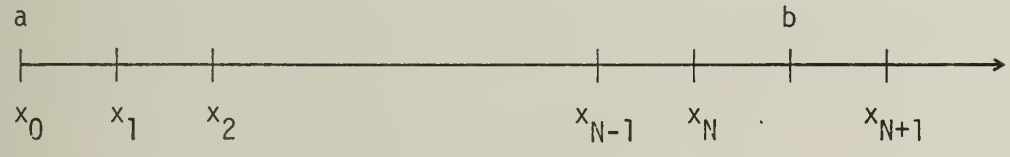
for $$g = \begin{pmatrix} g^1 \\ g(x),\ a \leq x \leq \bar{b} \\ g^2 \end{pmatrix} \in E^0$$

$$F(z) = \begin{pmatrix} z(a) - \alpha \\ z''(x) - f(x, z(x)) \qquad a \leq x \leq \bar{b} \\ z'(b) - \beta \end{pmatrix}$$

2. Discretization

Introduce the grid $x_i = a + i\,h$, $i = 0, 1, \ldots, N+1$,

$h = 2(b - a)/(2N + 1)$ i.e.



and discretize the problem according to

$$\frac{\xi_{i+1} - 2\xi_i + \xi_{i-1}}{h^2} - f(x_i, \xi_i) = 0 \qquad i = 1, 2, \ldots, N$$

$$\xi_0 - \alpha = 0$$

$$\frac{\xi_{N+1} - \xi_N}{h} - \beta = 0$$

In the operator formalism this means

define $$\Delta_h : E \to E_h \quad ; \quad E_h = \mathbb{R}^{N+2}$$

$$\Delta_h z = [z(x_0), z(x_1), \ldots, z(x_{N+1})]$$

$$\Delta_h^0 : E^0 \to E_h^0 \quad ; \quad E_h^0 = \mathbb{R} \times \mathbb{R}^N \times \mathbb{R}$$

$$\Delta_h^0 g = \begin{pmatrix} g^1 \\ g(x_i) \qquad i = 1, 2, \ldots, N \\ g^2 \end{pmatrix}$$

for

$$g = \begin{pmatrix} g^1 \\ g(x) \qquad a \leq x \leq \bar{b} \\ g^2 \end{pmatrix} \in E^0$$

and use the norms

$$||\xi||_{E_h} = \max_{0 \leq i \leq N+1} |\xi_i|$$

$$||\gamma||_{E_h^0} = |\gamma^1| + |\gamma^2| + \max_{1 \leq i \leq N} |\gamma_i|$$

for

$$\gamma = \begin{pmatrix} \gamma^1 \\ \gamma_i \qquad i = 1, 2, \ldots, N \\ \gamma^2 \end{pmatrix} \in E_h^0$$

Then define

$$\phi_h: \; E_h \to E_h^0$$

$$\phi_h(\xi) = \begin{pmatrix} \xi_0 - \alpha \\ \dfrac{\xi_{i+1} - 2\xi_i + \xi_{i-1}}{h^2} - f(x_i, \xi_i); \qquad i = 1, 2, \ldots, N \\ \dfrac{\xi_{N+1} - \xi_N}{h} - \beta \end{pmatrix}$$

Note that, by Taylor expansions we get

$$\frac{z(x_{i+1}) - 2z(x_i) + z(x_{i-1})}{h^2} - f(x_i, z(x_i)) = z''(x_i) - f(x_i, z(x_i))$$

$$+ \sum_{j=1}^{s-1} \frac{2}{(2j+2)!} z^{(2j+2)}(x_i)\, h^{2j}$$

$$+ O(h^{2s-1})$$

and

$$\frac{z(x_{N+1}) - z(x_N)}{h} - \beta = z'(b) - \beta + \sum_{j=1}^{s-1} \frac{2}{(2j+1)!} z^{(2j+1)}(b)(\tfrac{1}{2})^{2j+1}\, h^{2j}$$

$$+ O(h^{2s-1})$$

for $z \in E$. If we define $f_v\colon\ E \to E^0$, $v = 2, 3, \ldots, 2s-2$,

$$f_v(z) = 0 \qquad v \text{ odd}$$

$$f_v(z) = \begin{pmatrix} 0 \\ \frac{2}{(v+2)!} z^{(v+2)}(x) \qquad a \leq x \leq \bar{b} \\ \frac{2}{(v+1)!} z^{(v+1)}(b)(\tfrac{1}{2})^{v+1} \end{pmatrix} \qquad v \text{ even}$$

we have

$$\phi_h(\Delta_h z) = \Delta_h^0\{F(z) + \sum_{v=2}^{2s-2} f_v(z)\, h^v\} + O(h^{2s-1})$$

3. Perturbation

Define

$$D_i: \; E_h \to \mathbb{R} \qquad i = 1, 2, \ldots, N$$

$$D_i(\xi) = \begin{cases} \dfrac{50\xi_0 - 75\xi_1 - 20\xi_2 + 70\xi_3 - 30\xi_4 + 5\xi_5}{60h^2} & \text{for } i = 1 \\ \dfrac{-5\xi_{i-2} + 80\xi_{i-1} - 150\xi_i + 80\xi_{i+1} - 5\xi_{i+2}}{60h^2} & \text{for } i = 2, \ldots, N-1 \\ \dfrac{5\xi_{N-4} - 30\xi_{N-3} + 70\xi_{N-2} - 20\xi_{N-1} - 75\xi_N + 50\xi_{N+1}}{60h^2} & \text{for } i = N \end{cases}$$

and $R_b: \; E_h \to \mathbb{R}$

$$R_b(\xi) = \frac{-\xi_{N-3} + 5\xi_{N-2} - 9\xi_{N-1} - 17\xi_N + 22\xi_{N+1}}{24h}$$

Note that, by Taylor expansions we get

$$D_i(\Delta_h z) = z''(x_i) + O(h^4) \qquad i = 1, 2, \ldots, N$$

and

$$R_b(\Delta_h z) = z'(b) + O(h^4)$$

for $z \in E$.

Now define

$$\phi_h^E(\xi) = \begin{pmatrix} \xi_0 - \alpha \\ D_i(\xi) - f(x_i, \xi_i) \qquad i = 1, 2, \ldots, N \\ R_b(\xi) - \beta \end{pmatrix}$$

then for $z \in E$

$$\phi_h^E(\Delta_h z) = \Delta_h^0\{F(z)\} + O(h^4)$$

4. As a further exercise we examine some of the conditions of theorem 1

$$||f_j(y^1) - f_j(y^2)||_{E^0} = ||\begin{pmatrix} 0 \\ \frac{2}{(j+2)!}(y^{1(j+2)} - y^{2(j+2)}) \\ \frac{2}{(j+1)!}(y^{1(j+1)}(b) - y^{2(j+1)}(b))\cdot(\frac{1}{2})^{j+1} \end{pmatrix}||_{E_0}$$

$$= 2\cdot(\tfrac{1}{2})^{j+1}\,\frac{|y^{1(j+1)}(b) - y^{2(j+1)}(b)|}{(j+1)!}$$

$$+ 2\,\max_{a\leq x\leq b}\frac{|y^{1(j+2)}(x) - y^{2(j+2)}(x)|}{(j+2)!}$$

$$\leq 2\cdot(\tfrac{1}{2})^{j+1}\,||y^1 - y^2||_E + 2\cdot||y^1 - y^2||_E$$

$$\leq 3\cdot||y^1 - y^2||_E$$

for any $y^1, y^2 \in E$, $j = 2, 4, \ldots, 2s - 2$.

Further $$||f_j(y^1) - f_j(y^2)||_{E^0} = 0$$

for $j = 3, 5, \ldots, 2s - 3$

so condition d) of theorem 1 is satisfied.

$$||\phi_h(\xi^1) - \phi_h(\xi^2)||_{E_h^0}$$

$$= \left|\left|\begin{pmatrix} \xi_0^1 - \xi_0^2 \\ \dfrac{\xi_{i+1}^1 - \xi_{i+1}^2 - 2(\xi_i^1 - \xi_i^2) + \xi_{i-1}^1 - \xi_{i-1}^2}{h^2} - (f(x_i, \xi_i^1) - f(x_i, \xi_i^2)) \\ \dfrac{\xi_{N+1}^1 - \xi_{N+1}^2 - (\xi_N^1 - \xi_N^2)}{h} \end{pmatrix}\right|\right|_{E_h^0}$$

$$= |\xi_0^1 - \xi_0^2| + \frac{|\xi_{N+1}^1 - \xi_{N+1}^2 - (\xi_N^1 - \xi_N^2)|}{h}$$

$$+ \max_{1 \le i \le N} \left|\frac{\xi_{i+1}^1 - \xi_{i+1}^2 - 2(\xi_i^1 - \xi_i^2) + \xi_{i-1}^1 - \xi_{i-1}^2}{h^2} - f_y(x_i, \bar{\xi}_i)(\xi_i^1 - \xi_i^2)\right|$$

$$\le ||\xi^1 - \xi^2||_{E_h} + \frac{2}{h}||\xi^1 - \xi^2||_{E_h} + \frac{4}{h^2}||\xi^1 - \xi^2||_{E_h} + M_y||\xi^1 - \xi^2||_{E_h}$$

$$\le C \cdot ||\xi^1 - \xi^2||_{E_h} \cdot h^{-2}$$

where

$$C = h_0^2(1 + M_y) + 2h_0 + 4$$

$$\bar{\xi}_i \in \text{int}[\xi_i^1, \xi_i^2]$$

$$M_y = \max_{\substack{a \le x \le b \\ |z - y(x)| \le d}} |f_y(x, z(x)| \quad ; \quad ||\xi^i - \Delta_h y||_{E_h} \le d \qquad i = 1, 2$$

Analogously we get

$$||\phi_h^E(\xi^1) - \phi_h^E(\xi^2)||_{E_h^0} \le C_E \cdot ||\xi^1 - \xi^2||_{E_h} \cdot h^{-2}$$

where

$$C_E = h_0^2(1 + M_y) + \frac{54}{24} h_0 + \frac{320}{60}$$

and M_y, ξ^1, ξ^2 are as above. Thus condition e) is satisfied.

Theorem 3

Let y be the unique solution of $F(y) = 0$ and

a) the global discretization error $\eta - \Delta_h y$ of (2.3) have an asymptotic expansion

$$\eta - \Delta_h y = \Delta_h\{\sum_{j=p}^{M} h^j e_j\} + \delta^0(h)$$

with $M \geq \mu(p + k)$ and $||\delta^0|| = O(h^{M+1})$

b) the expansions (2.2) and (2.10) hold with $M \geq \mu(p + k)$

c) the operator ϕ_h be stable at $\Delta_h y$ in the sense of (2.6)

d) the operators F, f_j and $f_{i,j}$ have the following differentiability properties:

F: $[M/p] + 1$ times Frechet differentiable

f_j: $[(M - j)/p] + 1$ times Frechet differentiable

$j = p, p + 1, \ldots, M$

f_{vj}: $[(M_v - j)/(vp)] + 1$ times Frechet differentiable

$j = (v + 1)\cdot p, (v + 1)\cdot p + 1, \ldots, M_v$

$v = 1, 2, \ldots, \mu - 1$

Here [expression] stands for the integer part of the expression within the square bracket.

e) there exist constants d, C and C_v, $v = 1, 2, \ldots, \mu - 1$ such that uniformly for all $h \in H$

$$||\phi_h(\xi^1) - \phi_h(\xi^2)|| \leq C||\xi^1 - \xi^2||\cdot h^{-k}$$

$$||\phi_{h,v}(\xi^1) - \phi_{h,v}(\xi^2)|| \leq C_v||\xi^1 - \xi^2||\cdot h^{-k}$$

for all $\xi^i \in E_h$, $i = 1, 2$ such that

$$||\xi^i - \Delta_h y||_{E_h} \leq d$$

f) $[F'(y)]^{-1}$ exist

g) it be possible to define e_{vj}, $j = (v + 1)\, p, \ldots, M_v$, $v = 1, 2, \ldots, \mu - 1$, $M_v = M - k\cdot v$ according to

$$F'(y)\, e_{vj} = g_{vj}$$

where g_{vj} are defined in equation (2.18) below; then the global discretization error $\eta^i - \Delta_h y$ of (2.11) has an asymptotic expansion of the form

$$\eta^i - \Delta_h y = \Delta_h\{\sum_{j=(i+1)\cdot p}^{M_i} e_{ij}\, h^j\} + \delta^i(h)$$

where $||\delta^i|| = O(h^{M_i+1})$; $M_i = M - ik$, $i = 1, 2, \ldots, \mu - 1$

Proof

Introduce the family of operators $\psi_{h,i}$: $E_h \to E_h^0$, $i = 1, 2, \ldots, \mu - 1$ by

$$\psi_{h,i}(\xi) = \phi_h(\xi) + \sum_{j=1}^{i} \phi_{h,j}(\eta^{j-1}) \tag{2.13}$$

Then η^i, $i = 1, 2, \ldots, \mu - 1$ are the solutions of

$$\psi_{h,i}(\eta^i) = 0 \tag{2.14}$$

Note that as

$$0 = \psi_{h,i-1}(\eta^{i-1}) = \phi_h(\eta^{i-1}) + \sum_{j=1}^{i-1} \phi_{h,j}(\eta^{j-1})$$

we have

$$0 = \psi_{h,i}(\eta^{i}) = \phi_h(\eta^{i}) - \phi_h(\eta^{i-1}) + \phi_{h,i}(\eta^{i-1})$$

i.e.

$$\phi_h(\eta^{i}) = \phi_h(\eta^{i-1}) - \phi_{h,i}(\eta^{i-1})$$

Further note that as $\psi_{h,i}$ and ϕ_h only differ by an additive constant the stability of $\psi_{h,i}$ is implied by the stability of ϕ_h.

We will prove the theorem by induction on i, so assume that

$$\eta^{i-1} - \Delta_h y = \Delta_h\{\sum_{j=i\cdot p}^{M_{i-1}} e_{i-1\ j}\ h^j\} + \delta^{i-1}(h) \tag{2.15}$$

with $||\delta^{i-1}|| = O(h^{M_{i-1}+1})$ and $M_i = M - i\cdot k$.

Introduce the notation

$$z^s = \sum_{j=(s+1)\cdot p}^{M_s} e_{sj}\ h^j \qquad s = 0, 1, \ldots, \mu - 1 \tag{2.16}$$

From (2.14), (2.13), (2.2) and (2.10) we get

$$||\psi_{h,i}(\eta^{i}) - \psi_{h,i}(\Delta_h\{y + z^i\})|| = O(h^p)$$

so for sufficiently small h we have

$$||\psi_{h,i}(\eta^{i}) - \psi_{h,i}(\Delta_h\{y + z^i\})|| < r$$

Hence from the stability of $\psi_{h,i}$ we get for $h \in H$

$$||\delta^{i}|| = ||\eta^{i} - \Delta_h z^i - \Delta_h y|| = ||\eta^{i} - \Delta_h\{y + z^i\}|| \tag{2.17}$$

$$\leq S||\psi_{h,i}(\eta^{i}) - \psi_{h,i}(\Delta_h\{y + z^i\})||$$

To prove that $||\delta^i|| = O(h^{M_i+1})$ if $e_{i,j}$ are defined as in g) we note that

$$\psi_{h,i}(\eta^i) - \psi_{h,i}(\Delta_h\{y + z^i\}) = \phi_h(\eta^i) - \phi_h(\Delta_h\{y + z^i\})$$

$$= \phi_h(\eta^{i-1}) - \phi_{h,i}(\eta^{i-1}) - \phi_h(\Delta_h\{y + z^i\})$$

$$= \phi_h(\Delta_h\{y + z^{i-1}\}) - \phi_{h,i}(\Delta_h\{y + z^{i-1}\}) - \phi_h(\Delta_h\{y + z^i\}) + O(h^{M_{i-1}+1-k})$$

$$= \Delta_h^0\{F(y + z^{i-1}) + \sum_{j=p}^{M_i} f_j(y + z^{i-1})\, h^j - F(y + z^{i-1})$$

$$- \sum_{j=(i+1)\cdot p}^{M_i} f_{i,j}(y + z^{i-1})\, h^j - F(y + z^i) - \sum_{j=p}^{M_i} f_j(y + z^i)\, h^j\} + O(h^{M_i+1})$$

$$= \Delta_h^0\{- F(y + z^i) + \sum_{j=p}^{M_i} [f_j(y + z^{i-1}) - f_j(y + z^i)]\, h^j$$

$$- \sum_{j=p(i+1)}^{M_i} f_{i,j}(y + z^{i-1})\, h^j\} + O(h^{M_i+1})$$

$$= - \Delta_h^0\{F(y) + \sum_{r=1}^{M_i} \frac{1}{r!} F^{(r)}(y)(z^i)^r - \sum_{j=p}^{M_i} \Big(\sum_{r=1} \frac{1}{r!} f_j^{(r)}(y)[(z^{i-1})^r - (z^i)^r]\Big)\, h^j$$

$$+ \sum_{j=p(i+1)}^{M_i} [\sum_{r=0} \frac{1}{r!} f_{i,j}^{(r)}(y)(z^{i-1})^r]\, h^j\} + O(h^{M_i+1})$$

The upper summation limits in the two sums over r above are chosen such that all relevant terms are included.

Insert the expressions for z^i and z^{i-1} and collect terms of equal powers of h, then

$$\psi_{h,i}(\eta^i) - \psi_{h,i}(\Delta_h\{y + z^i\}) = - \Delta_h^0\{ \sum_{j=(i+1)p}^{M_i} (F'(y)\, e_{i\,j} - g_{i\,j})\, h^j\}$$

where $g_{i\,j}$ is independent of $e_{i\,j}$ and h and can be constructed from

$$(2.18)\qquad \sum_{j=(i+1)p}^{M_i} g_{ij}\,h^j = -\sum_{r=2}^{M_i}\frac{1}{r!}F^{(r)}(y)\Big(\sum_{\ell=(i+1)p}^{M_i} e_{i\ell}\,h^\ell\Big)^r$$

$$+\sum_{m=p}^{M_i}\Big(\sum_{r=1}\frac{1}{r!}f_m^{(r)}(y)\Big[\Big(\sum_{\ell=ip}^{M_{i-1}} e_{i-1\ell}\,h^\ell\Big)^r$$

$$-\sum_{\ell=(i+1)p}^{M_i} e_{i\ell}\,h^\ell\Big)^r\Big]\Big)h^m$$

$$-\sum_{m=(i+1)\cdot p}^{M_i}\Big[\sum_{r=0}\frac{1}{r!}f_{im}^{(r)}(y)\Big(\sum_{\ell=ip}^{M_{i-1}} e_{i-1\ell}\,h^\ell\Big)^r\Big]\cdot h^m$$

$$+O(h^{M_{i+1}})$$

Now, if e_{ij}, $j = (i+1)p, \ldots, M_i$ are the solutions of

$$F'(y)e_{ij} - g_{ij} = 0$$

then

$$\psi_{h,i}(\eta^i) - \psi_{h,i}(\Delta_h\{y + z^i\}) = O(h^{M_{i+1}})$$

and thus from (2.17)

$$||\delta^i|| = O(h^{M_{i+1}})$$

But from condition a) the assumption of induction is true for i = 0 with $\eta^0 = \eta$, thus the theorem is proved. Q.E.D.

Theorem 4

Let y be the unique solution of F(y) = 0 and let

a) all the conditions, except b), of theorem 3 hold

b) ϕ_h, ϕ_h^E be [M/p] + 1 times continuously Frechet differentiable

c) the inequalities below hold for v = 0, 1, ..., [M/p]

$$\phi_h^{(v)}(\Delta_h\, y) = \Delta_h^0\{F^{(v)}(y) + \sum_{j=p}^{M-v\cdot p} f_j^{(v)}(y)\, h^j\} + O(h^{M+1-vp})$$

$$\phi_{h,i}^{(v)}(\Delta_h\, y) = \Delta_h^0\{F^{(v)}(y) + \sum_{j=(i+1)\cdot p}^{M-vp} f_{i,j}^{(v)}(y)\, h^j\} + O(h^{M+1-vp})$$

for $i = 1, 2, \ldots, \mu - 1$

then the global discretization error $\eta^i - \Delta_h\, y$ of (2.11) has an asymptotic expansion of the form

$$\eta^i - \Delta_h\, y = \Delta_h\{\sum_{j=(i+1)p}^{M_i} e_{ij}\, h^j\} + \delta^i(h)$$

where

$$||\delta^i|| = O(h^{M_{i+1}}) \quad ; \quad M_i = M - i\cdot k$$

$i = 1, 2, \ldots, \mu - 1$

The proof of this theorem is analogous to the proof of theorem 3, cf. the proofs of theorem 1 and 2.

Note 1 These theorems are the basis for iterative improvement of the numerical solution of an operator equation. When the conditions of the theorem are satisfied at least $\mu - 1$ improvements can be made. At no extra expense an estimate of the global discretization error of the solution η^i can be obtained as $\eta^i - \eta^{i-1}$.

Note 2 The condition on M in the expansion (2.10) can be relaxed somewhat, and could be made dependent on i so M would decrease with i.

Note 3 Other combinations of the conditions on ϕ_h and $\phi_{h,v}$ of theorems 3 and 4 may be used. We can, e.g., use the conditions of b) and c) on $\phi_{h,v}$ from theorem 4 and the conditions on ϕ_h of d) from theorem 3.

Note 4 The main problem of applying this theorem is of course the construction of the operators $\phi_{h,r}$. Several examples of such operators will be given in section 4. To illustrate a major difficulty and resolve it we will again consider the two point boundary value problem

$$y'' = f(x, y)$$

$$y(a) = \alpha \qquad y'(b) = \beta$$

of note 5 to theorem 2. We cannot use the operator ϕ_h^E as $\phi_{h,1}$ because

$$D_1(\Delta_h z) = z''(x_1) + c_1 \cdot z^{(6)}(x_1) \cdot h^4$$

$$D_i(\Delta_h z) = z''(x_i) + c_0 \cdot z^{(6)}(x_i) \cdot h^4 \qquad c_1 \neq c_0 \neq c_2$$

$$D_N(\Delta_h z) = z''(x_N) + c_2 \cdot z^{(6)}(x_N) \cdot h^4$$

so the expansion (2.10) is not valid. This is caused by the use of approximation formulas for the two points closest to the boundary that are different from the formulas in the interior of the interval. However, if we fix the number of iterative improvements that we want to make to $(\mu - 1)$ we can construct operators D_i: $E_h \to \mathbb{R}$, $i = 1, 2, \ldots, N$

$$D_i(\xi) = \sum_{v=0}^{N+1} \alpha_{iv}\, \xi_v$$

(some $\alpha_{i\,v}$ may be zero) such that

$$D_i(\Delta_h z) = z''(x_i) + O(h^{\mu(p+k)})$$

Analogously construct operators R_b: $E_h \to \mathbb{R}$ such that

$$R_b(\Delta_h z) = z'(b) + O(h^{\mu(p+k)})$$

Define

$$\phi_{h,r}(\xi) = \begin{pmatrix} \xi_0 - \alpha \\ D_i(\xi) - f(x_i, \xi_i) \qquad i = 1, 2, \ldots, N \\ R_b(\xi) - \beta \end{pmatrix}$$

$r = 1, 2, \ldots, \mu - 1$

then

$$\phi_{h,r}(\Delta_h z) = \Delta_h^0\{F(z)\} + O(h^{\mu(p+k)})$$

Hence all $\phi_{h,r}$, $r = 1, 2, \ldots, \mu - 1$ are identical and they are consistent of order $\mu(p + k)$ with F. From the formulas it is obvious that we do not need to use this trick for the approximations to $z'(b)$ but can use operators $R_{b,r}: E_h \to \mathbb{R}$ such that

$$R_{b,r}(\Delta_h z) = z'(b) + \sum_{v=(r+1)\cdot p}^{M} h^v c_{rv} + O(h^{M+1})$$

in the definition of $\phi_{h,r}$.

3. Approximation of linear functionals

To construct the perturbation operators ϕ_h^E and $\phi_{h,i}$, $i = 1, 2, \ldots$ in the applications of section 4 we need in many cases formulas that approximate the value and the value of the derivatives of a function z for a given argument by linear combinations of the components of $\Delta_h z$.

Such formulas have been examined in Ballester, Pereyra (1967), Bjorck, Pereyra (1970), Galimberti, Pereyra (1970), (1971), and Pereyra (1973), so we simply refer to those papers and introduce some notation that will be useful in section 4.

Let

$$E = C^S(a, b) \quad , \quad E_h = \mathbb{R}^{P+1}$$

Define

$$\Delta_h : \; E \to E_h$$

$$\Delta_h z = [z(x_0), z(x_1), \ldots, z(x_P)]$$

$$x_j \in [a, b] \qquad j = 0, 1, \ldots, P$$

for $z \in E$.

Define the linear operators (depending on h)

$$D_x^{v,m} : \; E_h \to \mathbb{R} \qquad \begin{array}{l} x \in [a, b] \\ v = 0, 1, \ldots, P \\ m = 1, 2, \ldots, P + 1 - v \end{array}$$

according to

$$D_x^{v,m}(\xi) = \sum_{r=0}^{P} \alpha_{v,m,r}(x)\, \xi_r$$

(most of the $\alpha_{v,m,r}(x)$ may be zero) such that

$$D_x^{v,m}(\Delta_h z) = z^{(v)}(x) + \sum_{j=m}^{S} g_j(z)(x)\, h^j + O(h^{S+1})$$

for $z \in E$.

Note: As P is finite the order of consistency m of the operator $D_x^{v,m}$ with $z^{(v)}(x)$ cannot exceed P + 1 - v.

When v, m and x are fixed the constants $\alpha_{v,m,r}(x)$; r = 0, 1, ..., P can be obtained as the solution of a Vandermonde system of linear equations. In Bjorck, Pereyra (1970) an efficient algorithm for the numerical solution of such linear systems is given.

For a Vandermonde system with n unknowns this algorithm requires approximately $3n^2$ operations. To get $D_x^{v,m}$ consistent of order m with $z^{(v)}(x)$ we only need to have m + v of the coefficients $\alpha_{v,m,r} \neq 0$, i.e. the Vandermonde system we have to solve has only m + v unknowns. Further most of the operators $D_x^{v,m}$ needed have $x = x_i$, where x_i is a gridpoint. Once we have found the operator $D_x^{v,m}$ for one gridpoint we can easily get the operators $D_x^{v,m}$ for most of the other gridpoints (if the gridpoints are equidistant) without solving a system of linear equations. In conclusion, the amount of work needed to find these operators is in general negligible compared to the amount of work involved in solving the operator equations

$$\phi_h(\eta) = 0 \text{ and } \phi_h(\eta^E) + \phi_h^E(\eta) = 0$$

or

$$\phi_h(\eta^i) + \sum_{v=1}^{i} \phi_{h,i}(\eta^{i-1}) = 0 \quad , \quad i = 1, 2, \ldots$$

For some examples of operators of this type see note 6 after theorems 1 and 2.

For functions of several variables the formulas above can be used for each of the variables or one can use more compact approximation formulas taking linear combinations of the function values at all the gridpoints.

To distinguish between the different partial derivatives we adjoin to D the variable that we differentiate with respect to, e.g.

$$Dx^{v,m}_{x,y} \text{ is consistent of order m with } \left(\frac{\partial^v}{\partial x^v}\right)(x, y)$$

$$DT^{v,m}_{x,y,z,t} \text{ is consistent of order m with } \left(\frac{\partial^v}{\partial t^v}\right)(x, y, z, t)$$

4. Applications

The main emphasis of these applications is the construction of the perturbation operators ϕ_h^E and $\phi_{h,i}$, $i = 1, 2, \ldots$. The operators ϕ_h^E all correspond to well known discretizations from the literature. We do not claim that the methods chosen are the best possible for the actual problems; rather we have tried to use methods that are fairly well known and in some cases widely used.

Although the examples of this section are mainly given as illustrations of how to apply the general ideas of section 2 to different classes of problems, we believe that the algorithms for iterative improvement of elliptic partial differential equations and for iterative improvement of integral equations compare very favorably with existing algorithms for these classes of problems.

The questions of the smoothness of the expansions of the global discretization errors for the basic discretizations have not been examined in detail. For the different applications, wherever possible, reference to known results on such expansions are given, while in other cases we discuss very briefly the possible existence of such expansions. These questions need further study.

In the theorems of section 2 we make some assumptions on the perturbation operators ϕ_h^E and $\phi_{h,i}$, $i = 1, 2, \ldots$. The validity of these assumptions can easily be checked in all the applications.

In future studies on each of the applications of interest to us a more detailed analysis will be given.

Some numerical results are given in most of the subsections below.

4.1 Initial value problems for ordinary differential equations

Consider the scalar differential equation

$$y' = f(y) \quad ; \quad y(0) = \alpha \qquad x \in [0, T] \qquad f \in C^S(\mathbb{R})$$

The use of an autonomous problem is only for convenience in notation, as is the restriction to a scalar differential equation. All the results can easily be generalized to non-autonomous systems of ordinary differential equations. The existence of smooth asymptotic error expansions for discretization methods for initial value problems for ordinary differential equations is discussed in Stetter (1965), (1973).

In our operator formalism the problem is

$$F: \; E \to E^0$$

$$E = C^S(0, T) \qquad ||z||_E = \max_{0 \le x \le T} \sum_{v=0}^{S} \frac{|z^{(v)}(x)|}{v!}$$

$$E^0 = \mathbb{R} \times C(0, T) \qquad ||g||_{E^0} = |\gamma^1| + \max_{0 \le x \le T} |g(x)|$$

$$\text{for } g = \begin{pmatrix} \gamma^1 & \\ g(x) & 0 \le x \le T \end{pmatrix}.$$

$$F(z) = \begin{pmatrix} z(0) - \alpha & \\ z' - f(z) & 0 \le x \le T \end{pmatrix}$$

Consider Euler's method

$$y_0 = \alpha$$

$$y_{i+1} - y_i = h \cdot f(y_i) \qquad i = 0, 1, \ldots, N - 1 \qquad h = T/N$$

This simple method is chosen for illustrative purposes. Although the analysis is simple for this method the techniques used to construct the perturbation operators are applicable to more complicated methods and more complicated problems.

Euler's method in our operator formalism reads:

$$\Delta_h: \; E \to E_h \quad ; \quad E_h = \mathbb{R}^{N+1} \qquad ||\xi||_{E_h} = \max_{0 \leq i \leq N} |\xi_i|$$

$$\Delta_h z = (z(x_0), z(x_1), \ldots, z(x_N))$$

$$x_i = i \cdot h \quad ; \quad i = 0, 1, \ldots, N \quad ; \quad h = T/N$$

Let

$$\Delta_h^0: \; E^0 \to E_h^0 \quad ; \quad E_h^0 = \mathbb{R} \times \mathbb{R}^N$$

$$\Delta_h^0 g = \begin{pmatrix} g^1 & \\ & \\ g(x_i) & i = 0, 1, \ldots, N-1 \end{pmatrix} \quad \text{for } g = \begin{pmatrix} g^1 & \\ & \\ g(x) & 0 \leq x \leq T \end{pmatrix}$$

$$||\gamma||_{E_h^0} = |\gamma^1| + \max_{0 \leq i \leq N-1} |\gamma_i|$$

and

$$\phi_h: \; E_h \to E_h^0$$

$$\phi_h(\xi) = \begin{pmatrix} \xi_0 - \alpha & \\ & \\ \dfrac{\xi_{i+1} - \xi_i}{h} - f(\xi_i) & i = 0, 1, \ldots, N-1 \end{pmatrix}$$

Note that

$$\frac{z(x_{i+1}) - z(x_i)}{h} - f(z(x_i)) = z'(x_i) - f(z(x_i)) + \sum_{j=1}^{S-1} \frac{z^{(j+1)}(x_i)}{(j+1)!} h^j + O(h^S)$$

for any $z \in E$

so

$$\phi_h(\Delta_h z) = \Delta_h^0 \{F(z) + \sum_{j=1}^{S-1} f_j(z) \, h^j\} + O(h^S)$$

where

$$f_j(z) = \begin{pmatrix} 0 \\ \\ \dfrac{z^{(j+1)}(x)}{(j+1)!} & 0 \leq x \leq T \end{pmatrix}$$

Now construct ϕ_h^E and $\phi_{h,v}$, $v = 1, 2, \ldots$. We will give some different operators ϕ_h^E, $\phi_{h,v}$ in order to illustrate some useful techniques for construction of these operators.

1.

$$\phi_h^E(\xi) = \begin{pmatrix} \xi_0 - \alpha \\ \\ \dfrac{-\xi_2 + 4\xi_1 - 3\xi_0}{2h} - f(\xi_0) \\ \\ \dfrac{\xi_{i+1} - \xi_{i-1}}{2h} - f(\xi_i) & i = 1, 2, \ldots, N-1 \end{pmatrix}$$

Note that

$$\frac{-z(x_2) + 4z(x_1) - 3z(x_0)}{2h} - f(z(x_0)) = z'(x_0) - f(z(x_0)) + \sum_{j=2}^{S-1} \frac{4 - 2^{j+1}}{2(j+1)!} z^{(j+1)}(x_0) \cdot h^j + O(h^S)$$

and

$$\frac{z(x_{i+1}) - z(x_{i-1})}{2h} - f(z(x_i)) = z'(x_i) - f(z(x_i)) + \sum_{\ell=1}^{(S-1)/2} \frac{1}{(2\ell+1)!} z^{(2\ell+1)}(x_i) \cdot h^{2\ell} + O(h^S)$$

for $i = 1, 2, \ldots, N-1$

Both the expansions are valid for any $z \in E$. Hence, for any $z \in E$

$$\phi_h^E(\Delta_h z) = \Delta_h^0\{F(z)\} + O(h^2)$$

but we do not have

$$\phi_h^E(\Delta_h z) = \Delta_h^0\{F(z) + \sum_{j=2}^{S-1} f_j(z)\, h^j\} + O(h^S)$$

because of the different expansions for x_0 and the rest of the gridpoints!

2. By extending the operators so an approximation to the solution at $x_{-1} = -h$ is computed by the formula

$$\frac{y_0 - y_{-1}}{h} - f(y_{-1}) = 0$$

we can define

$$\phi_h^E(\xi) = \begin{pmatrix} \xi_0 - \alpha \\ \\ \dfrac{\xi_{i+1} - \xi_{i-1}}{2h} - f(\xi_i) \qquad i = 0, 1, \ldots, N-1 \end{pmatrix}$$

Alternatively we can expand the operators so that an approximation at $x_{N+1} = x_N + h$ is computed by the formula

$$\frac{y_{N+1} - y_N}{h} - f(y_N) = 0$$

and define

$$\phi_h^E(\xi) = \begin{pmatrix} \xi_0 - \alpha \\ \\ -\dfrac{\xi_{i+2} - 4\xi_{i+1} - 3\xi_i}{2h} - f(\xi_i) \qquad i = 0, 1, \ldots, N-1 \end{pmatrix}$$

In both cases we have, for any $z \in E$

$$\phi_h^E(\Delta_h z) = \Delta_h^0\{F(z) + \sum_{j=2}^{S-1} f_j(z)\, h^j\} + O(h^S)$$

The operators f_j will of course be different for the two different operators ϕ_h^E. By extending the numerical solution even further outside the interval of interest in this way one can easily construct operators $\phi_{h,v}$,

$v = 1, 2, \ldots$ such that

$$\phi_{h,v}(\Delta_h z) = \Delta_h^0\{F(z) + \sum_{j=(v+1)}^{S-1} f_{v,j}(z)\, h^j\} + O(h^S)$$

For these extended operators the spaces E, E^0, E_h and E_h^0 must be modified in order that the theorems of section 2 be applicable. Slight modifications of the theorems may also be necessary.

3.
$$\phi_{h,v}(\xi) = \begin{pmatrix} \xi_0 - \alpha \\ \\ D_{x_i}^{1,S}(\xi) - f(\xi_i) \qquad i = 0, 1, \ldots, N-1 \end{pmatrix}$$

$v = 1, 2, \ldots$

Here $D_{x_i}^{1,S}$ are operators of the type discussed in section 3. For any $z \in E$

$$\phi_{h,v}(\Delta_h z) = \Delta_h^0\{F(z)\} + O(h^S)$$

$v = 1, 2, \ldots$

4.
$$\phi_h^E(\xi) = \begin{pmatrix} \xi_0 - \alpha \\ \\ \dfrac{\xi_{i+1} - \xi_i}{h} - \dfrac{1}{2}[f(\xi_{i+1}) + f(\xi_i)] \qquad i = 0, 1, \ldots, N-1 \end{pmatrix}$$

Note that for $y \in E$ such that $y' = f(y)$

$$\frac{y(x_{i+1}) - y(x_i)}{h} - \frac{1}{2}[f(y(x_i)) + f(y(x_{i+1}))]$$

$$= y'(x_i) - f(y(x_i)) + \sum_{v=2}^{S} \frac{y^{(v)}(x_i)}{v!} h^{v-1} - \frac{1}{2} \sum_{r=1}^{S} f^{(r)}(y(x_i))\Big(\sum_{t=1}^{S} \frac{y^{(v)}(x_i)}{t!} h^t\Big)^r$$

$$+ O(h^S)$$

$$= y'(x_i) - f(y(x_i)) + \sum_{j=2}^{S-1} g_j(y(x_i)) \cdot h^j + O(h^S)$$

The absence of a term $g_1(y(x_i))h$ is due to the fact that for the elements y that we consider we have $y'' = f'(y)\, y'$!

Note that the expansion above is not valid for arbitrary $z \in E$, but only for $z \in E$ such that $z' = f(z)$. In this case we have to rely on theorem 2 while for the previous perturbation operators we can rely on both theorem 1 and theorem 2.

Note that for $y \in E \ni y' = f(y)$,

$$\phi_h^E(\Delta_h\, y) = \Delta_h^0\{F(y) + \sum_{j=2}^{S-1} f_j^E(y)\, h^j\} + O(h^S)$$

Further note that any method of second order or more can be used to construct the perturbation operator ϕ_h^E. The main advantages with our perturbation operator (which is based on the trapezoidal method) is that

1) no new values of $f(y)$ need to be computed to get the perturbation
2) the perturbed solution at x_i can be computed directly after the calculation of the unperturbed solution at x_i
3) the basic discretization formulas for ϕ_h and ϕ_h^E are both one-step formulas.

For general linear multistep methods

$$\sum_{v=0}^{k} \alpha_v y_{n+v} - h \sum_{v=0}^{k} \beta_v f_{n+v} = 0$$

one can e.g. use the following basic discretization formula for the perturbation operator

$$\sum_{v=0}^{k} \alpha_v^* y_{n+v} - h \sum_{v=0}^{k} \beta_v^* f_{n+v}$$

(We assume that the global discretization error for the solution obtained by the linear multistep method has a smooth error expansion; this may not be true in many cases!) Remember that the maximal order of a stable linear multistep method of stepnumber k is k + 1 when k is odd, and k + 2 when k is even (see e.g. Lambert (1973)). By choosing α_v^*, β_v^*, v = 0, 1, ..., k judiciously one can get

$$\sum_{v=0}^{k} \alpha_v^* y(x_{n+v}) - \sum_{v=0}^{k} \beta_v^* f(y(x_{n+v})) = y'(x_n) - f(y(x_n)) + O(h^{2k})$$

for $y \in E$ such that $y' = f(y)$.

One can of course also use perturbation operators of the type discussed in point 3 above. The basic discretization operator for ϕ_h^E here, however, has the same stepnumber k as the basic discretization operator for ϕ_h while the stepnumber for the basic discretization operators in point 3 may be considerably larger.

No study of the practical implementation of these ideas for initial value problems for ordinary differential equations has been undertaken yet.

5. As a further illustration we choose a different operator

$$\Delta_h^0 : \; E_h \to E_h^0 \quad ; \quad E_h^0 = R \times R^N$$

$$\Delta_h^0 g = \begin{pmatrix} g^1 \\ \\ g(x_i) \qquad i = 1, 2, \ldots, N \end{pmatrix} \quad \text{for } g = \begin{pmatrix} g^1 \\ \\ g(x) \qquad 0 \leq x \leq T \end{pmatrix}$$

$$||\gamma||_{E_h^0} = |\gamma^1| + \max_{1 \leq i \leq N} |\gamma_i|$$

The operator ϕ_h is the same as above, i.e.

$$\phi_h(\xi) = \begin{pmatrix} \xi_0 - \alpha \\ \\ \dfrac{\xi_{i+1} - \xi_i}{h} - f(\xi_i) \qquad i = 0, 1, \ldots, N-1 \end{pmatrix}$$

Note that for $i = 0, 1, \ldots, N-1$ and any $z \in E$

$$\frac{z(x_{i+1}) - z(x_i)}{h} - f(z(x_i)) = z'(x_{i+1}) - f(z(x_{i+1})) + \sum_{j=1}^{s-1} g_j(z(x_{i+1}))\, h^j + O(h^s)$$

Choose

$$\phi_h^E(\xi) = \begin{pmatrix} \xi_0 - \alpha \\ \\ \dfrac{\xi_{i+1} - \xi_{i-1}}{2h} - f(\xi_i) \qquad i = 1, 2, \ldots, N \end{pmatrix}$$

then

$$\phi_h^E(\Delta_h z) = \Delta_h^0 \{F(z) + \sum_{j=2}^{s-1} f_j(z)\, h^j\} + O(h^s)$$

Note 1 Some of the results of this section carries over to the case when x_i, $i = 0, 1, \ldots, N-1$ are not equidistant. E.g., the perturbations of point 3 and point 4 are useful in such cases.

Note 2 Consider again the scalar differential equation

$$y' - f(y) = 0 \qquad y(0) = \alpha \qquad ; \qquad x \in [0, T] \qquad ; \qquad f \in C^S(\mathbb{R})$$

Note that this problem is equivalent to the Volterra integral equation

$$y(x) - y(0) - \int_0^x f(y(s))\, ds = 0$$

Euler's method for the original differential equation can be viewed as an approximation to the integral equation, namely

$$y_{i+1} = y_i + h\, f(y_i) = y_{i-1} + h\, f(y_{i-1}) + h\, f(y_i) = \ldots$$

$$= \alpha + h \sum_{j=0}^{i} f(y_j) \qquad i = 0, 1, 2, \ldots$$

With appropriate definitions of the spaces E, E^0, E_h, E_h^0, the operators Δ_h, Δ_h^0 and the corresponding norms we have

$$F(z) = \left(z - \alpha - \int_0^x f(z(s))\, ds \quad , \quad 0 \leq x \leq T\right)$$

$$\phi_h(\xi) = \begin{pmatrix} \xi_0 - \alpha \\ \\ \xi_i - \alpha - h \sum_{j=0}^{i-1} f(\xi_j) \qquad ; \qquad i = 1, 2, \ldots, N \end{pmatrix}$$

For the actual computations we would of course use the recursion formula

$$\xi_0 = \alpha$$

$$\xi_i = \xi_{i-1} + h\, f(\xi_{i-1}) \qquad i = 1, 2, \ldots, N$$

Note that ϕ_h is consistent of order 1 with F, i.e. with the integral equation.

Define

$$\phi_h^E(\xi) = \begin{pmatrix} \xi_0 - \alpha \\ \\ \xi_i - \alpha - h \sum_{j=0}^{i} \alpha_{ij} f(\xi_j) \end{pmatrix} \quad , \quad i = 1, 2, \ldots, N$$

where

$$\alpha_{i0} = \alpha_{ii} = 1/2 \quad ; \quad \alpha_{ij} = 1 \quad j = 1, 2, \ldots, i - 1$$

For the actual computations we would use the formula

$$[\phi_h^E(\xi)]_0 = \xi_0 - \alpha$$

$$[\phi_h^E(\xi)]_i = [\phi_h^E(\xi)]_{i-1} + \xi_i - \xi_{i-1} - \frac{h}{2}[f(\xi_i) + f(\xi_{i-1})] \quad i = 1, 2, 3, \ldots, N$$

Note that ϕ_h^E is consistent of order 2 with F. Other perturbation operators ϕ_h^E could be used!

The advantage of this point of view is that for the maximum norms

$$||\gamma||_{E_h^0} = |\gamma^1| + \max_{0 \leq i \leq N} |\gamma_i|$$

where

$$\gamma = \begin{pmatrix} \gamma^1 \\ \\ \gamma_i \end{pmatrix} \quad i = 1(1)N \in E_h^0$$

and

$$||\xi||_{E_h} = \max_{0 \leq i \leq N} |\xi_i|$$

where

$$\xi = (\xi_i \quad , \quad i = 0(1)N) \in E_h$$

we have

$$||\phi_h(\xi^1) - \phi_h(\xi^2)||_{E_h^0} \leq C \cdot ||\xi^1 - \xi^2||_{E_h}$$

$$||\phi_h^E(\xi^1) - \phi_h^E(\xi^2)||_{E_h^0} \leq C_E \cdot ||\xi^1 - \xi^2||_{E_h}$$

i.e. $k = 0$ in condition e) of theorem 1, while for Euler's method directly we have $k = 1$ for the maximum norm. The same result, i.e. $k = 0$, may be obtained for Euler's method directly if we use Spijker's norm for E_h^0 (see Stetter (1973), section 2.2.4, p. 81-84) which in our case reads

$$||\gamma||_{E_h^0} = |\gamma^1| + h \max_{1 \leq v \leq N} \left| \sum_{j=1}^{v} \gamma_j \right|$$

for

$$\gamma = \begin{pmatrix} \gamma^1 \\ \gamma_i \qquad i = 1(1)N - 1 \end{pmatrix} \in E_h^0$$

Note the relation between Spijker's norm for E_h^0 corresponding to Euler's method and the maximum norm for E_h^0 corresponding to the integral equation formulation.

The reformulation of the original problem as an integro-differential equation may be a useful trick for other methods for initial value problems for ODEs and initial boundary value problems for PDEs.

The ideas presented in this note need further investigations and are included mainly as an illustration of a way to increase the range of problems and methods for which the theorems may be useful.

4.2 Two-point boundary value problems for ordinary differential equations

Consider

$$y'' = f(x, y, y') \qquad a - \varepsilon \leq x \leq b + \varepsilon$$

$$r_1(y(a), y'(a)) = 0 \qquad r_2(y(b), y'(b)) = 0$$

Assume that the functions f, r_1 and r_2 are such that the boundary value problem has a unique solution which is 2s times differentiable.

Define

$$F: \; E \to E^0$$

$$E = C^{2s}(a - \varepsilon, b + \varepsilon) \quad ; \quad ||z||_E = \max_x \sum_{v=0}^{2s} \frac{|z^{(v)}(x)|}{v!}$$

$$E^0 = R \times C(a, b) \times R$$

$$||g||_{E^0} = |\gamma^1| + |\gamma^2| + \max_x |g(x)|$$

for $$g = \begin{pmatrix} \gamma^1 \\ g(x) \qquad a - \varepsilon \leq x \leq b + \varepsilon \\ \gamma^2 \end{pmatrix} \in E^0$$

$$F(z) = \begin{pmatrix} r_1(z(a), z'(b)) \\ z'' - f(x, z, z') \qquad a - \varepsilon \leq x \leq b + \varepsilon \\ r_2(z(b), z'(b)) \end{pmatrix}$$

Introduce

$$\Delta_h: \; E \to E_h \quad ; \quad E_h = \mathbb{R}^{N+2}$$

$$\Delta_h z = [z(x_0), z(x_1), \ldots, z(x_{N+1})]$$

where $x_i = a - \frac{h}{2} + i \cdot h$, $i = 0, 1, \ldots, N + 1$; $h = \frac{b - a}{N}$

and $$\Delta_h^0: \; E^0 \to E_h^0 \quad ; \quad E_h^0 = \mathbb{R} \times \mathbb{R}^N \times \mathbb{R}$$

Define

$$\phi_h: \; E_h \to E_h^0$$

$$\phi_h(\xi) = \begin{pmatrix} r_1\left(\frac{\xi_1 + \xi_0}{2}, \frac{\xi_1 - \xi_0}{h}\right) \\ \frac{\xi_{i+1} - 2\xi_i + \xi_{i-1}}{h^2} - f\left(x_i, \xi_i, \frac{\xi_{i+1} - \xi_{i-1}}{2h}\right) \qquad i = 1, 2, \ldots, N \\ r_2\left(\frac{\xi_{N+1} + \xi_N}{2}, \frac{\xi_{N+1} - \xi_N}{h}\right) \end{pmatrix}$$

Assume that f, r_1 and r_2 are such that ϕ_h is stable. ϕ_h is consistent of order 2 with F.

The existence of smooth error expansion is discussed in Stetter (1965), Pereyra (1968).

Construction of perturbation operators

1. Direct approach.

Define $$\phi_h^E: \; E_h \to E_h^0$$

$$\phi_h^E(\xi) = \begin{pmatrix} r_1\left(D_a^{0,4}(\xi), D_a^{1,4}(\xi)\right) \\ D_{x_i}^{2,4}(\xi) - f\left(x_i, \xi_i, D_{x_i}^{1,4}(\xi)\right) \qquad i = 1, 2, \ldots, N \\ r_2\left(D_b^{0,4}(\xi), D_b^{1,4}(\xi)\right) \end{pmatrix}$$

where $D_x^{v,m}: \; E_h \to \mathbb{R}$ are operators of the kind discussed in section 3.

Define $\phi_{h,i}: E_h \to E_h^0 \qquad i = 1, 2, \ldots, \mu - 1$

$$\phi_{h,i}(\xi) = \begin{pmatrix} r_1(D_a^{0,2(i+1)}(\xi),\ D_a^{1,2(i+1)}(\xi)) \\ D_{x_i}^{2,2\mu}(\xi) - f(x_i, \xi_i, D_{x_i}^{1,2\mu}(\xi)) \qquad i = 1, 2, \ldots, N \\ r_2(D_b^{0,2(i+1)}(\xi),\ D_b^{1,2(i+1)}(\xi)) \end{pmatrix}$$

Note that for $\phi_{h,i}$ we have used the trick described in note 4 of theorem 4 in order to satisfy (2.10).

2. Use of Cowell's method.

If f is independent of y', i.e. $f(x, y, y') = g(x, y)$ define $\phi_h^E: E_h \to E_h^0$

$$\phi_h^E(\xi) = \begin{pmatrix} r_1(D_a^{0,4}(\xi),\ D_a^{1,4}(\xi)) \\ \dfrac{\xi_{i+1} - 2\xi_i + \xi_{i-1}}{h^2} - \dfrac{1}{12}\,(g(x_{i-1}, \xi_{i-1}) + 10g(x_i, \xi_i) \\ \qquad + g(x_{i+1}, \xi_{i+1})); \qquad i = 1, 2, \ldots, N \\ r_2(D_b^{0,4}(\xi),\ D_b^{1,4}(\xi)) \end{pmatrix}$$

In this case we rely on theorem 2 because $\phi_h^E(\Delta_h z) = \phi_h^0\{F(z)\} + O(h^4)$ only for $z \in E \ni z'' = g(x, z)$.

3. Extending the solution outside the interval $[a - \varepsilon, b + \varepsilon]$ (from an unpublished paper by H. Keller, V. Pereyra, communicated by V. Pereyra).

To be able to use symmetric and identical formulas to approximate $z^{(v)}(x_i)$, $v = 1, 2$ at all gridpoints x_i (including the gridpoints close to the boundary) one can extend the numerical solution to "artificial" gridpoints outside the interval (a, b) by the basic formula

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - f(x_n, y_n) = 0$$

Sufficiently many exterior points are introduced, so symmetrical and identical approximation formulas for the first and second derivative of any $z \in E$ can be used for all gridpoints needed. Now introduce $\phi_{h,i}$, $i = 1, 2, \ldots$

$$\phi_{h,i}(\xi) = \begin{pmatrix} r_1(D_a^{0,2(i+1)}(\xi), D_a^{1,2(i+1)}(\xi)) \\ D_{x_j}^{2,2(i+1)}(\xi) - f(x_j, \xi_j, D_{x_j}^{1,2(i+1)}(\xi)) \quad j = 1, 2, \ldots, N \\ r_2(D_b^{0,2(i+1)}(\xi), D_b^{1,2(i+1)}(\xi)) \end{pmatrix}$$

Then

$$\phi_{h,i}(\Delta_h z) = \Delta_h^0\{F(z) + \sum_{j=(i+1)\cdot p}^{2s-2} f_{ij}(z)\, h^j\} + O(h^{2s-1})$$

with a proper definition of Δ_h and Δ_h^0.

Numerical results

The following special cases of (1)

(A) $$y'' - f(x, y, y') = 0$$
$$y(a) = \alpha \qquad y(b) = \beta$$

(B) $$y'' - f(x, y, y') = 0$$
$$y(a) = \alpha \qquad y'(b) = \beta$$

were discretized by formulas analogous to those above. As the boundary conditions for these cases are simpler than in the general case, somewhat different discretizations are used here:

(A) $$x_i = a + i\cdot h \qquad i = 0, 1, \ldots, N \qquad h = \frac{b - a}{N}$$

$$\xi_0 - \alpha = 0$$

$$\frac{\xi_{i+1} - 2\xi_i + \xi_{i-1}}{h^2} - f(x_i, \xi_i, \frac{\xi_{i+1} - \xi_{i-1}}{2h}) = 0$$

$$i = 1, 2, \ldots, N - 1$$

$$\xi_N - \beta = 0$$

and

(B) $$x_i = a + i \cdot h \qquad i = 0, 1, \ldots, N \qquad h = \frac{2(b - a)}{2N - 1}$$

$$\xi_0 - \alpha = 0$$

$$\frac{\xi_{i+1} - 2\xi_i + \xi_{i-1}}{h^2} - f(x_i, \xi_i, \frac{\xi_{i+1} - \xi_{i-1}}{2h}) = 0$$

$$\frac{\xi_N - \xi_{N-1}}{h} - \beta = 0$$

The perturbation operators ϕ_h^E we use are

(A)
$$\phi_h^E(\xi) = \begin{pmatrix} \xi_0 - \alpha \\ D_{x_i}^{2,4}(\xi) - f(x_i, \xi_i, D_{x_i}^{1,4}(\xi)) \qquad i = 1, 2, \ldots, N - 1 \\ \xi_N - \beta \end{pmatrix}$$

(B)
$$\phi_h^E(\xi) = \begin{pmatrix} \xi_0 - \alpha \\ D_{x_i}^{2,4}(\xi) - f(x_i, \xi_i, D_{x_i}^{1,4}(\xi)) \qquad i = 1, 2, \ldots, N - 1 \\ D_b^{1,4}(\xi) - \beta \end{pmatrix}$$

The systems of non-linear equations obtained by the discretization were solved by Newton iteration and the initial approximation 0 was used.

The iterations were terminated when the last correction in the iterations was less than ε.

The perturbed problem was solved by a modified Newton iteration, where the LR-factorization of the final iteration matrix obtained in the solution of the unperturbed problem was used. As initial approximation the solution of the unperturbed problem was used and the iterations were carried on until an estimate of the iteration error was either less than one tenth of the estimated discretization error or less than ε. All quantities were measured in the maximum norm.

In the table below the entries of the column "Number of iterations" stand for the number of iterations for the unperturbed problem/number of iterations for the perturbed problem.

The following differential equations were solved:

(A1) $$y'' = \frac{1}{2}(y + x + 1)^3 \qquad y(0) = y(1) = 0$$

exact solution: $y(x) = 2/(2 - x) - x - 1$

(Ciarlet, Schultz and Varga (1967))

(A2) $$y'' = y^3 - \sin(x) * (1 + [\sin(x)]^2) \qquad y(0) = y(\pi) = 0$$

exact solution: $y(x) = \sin(x)$

(Pereyra (1968))

(A3) $$y'' = -1 - 0.49(y')^2 \qquad y(0) = y(1) = 0$$

exact solution: $y(x) = \frac{1}{0.49} \ln\left(\frac{\cos(0.7(x - 0.5))}{\cos(0.35)}\right)$

(Bellman, Kalaba (1965))

(B1) $$y'' = \frac{1}{2}(y + x + 1)^3 \qquad y(0) = 0 \qquad y'(1) = 1$$

exact solution: $y(x) = 2/(2 - x) - x - 1$

The problems were solved in single precision on an IBM 360/75 with $N = 10$ and $\varepsilon = 10^{-6}$.

Problem	||Actual error||	||Estimated error||	Number of iterations	||Solution||
A1	$7.3 \cdot 10^{-4}$	$7.1 \cdot 10^{-4}$	4/2	$1.7 \cdot 10^{-1}$
A2	$2.38 \cdot 10^{-3}$	$2.36 \cdot 10^{-3}$	6/2	1
A3	$1.059 \cdot 10^{-4}$	$1.060 \cdot 10^{-4}$	3/2	$1.3 \cdot 10^{-1}$
B1	$1.12 \cdot 10^{-3}$	$1.04 \cdot 10^{-3}$	4/2	$1.7 \cdot 10^{-1}$

Obviously one can get very good error estimates at a low cost for these problems. All the problems are very simple with fairly rapid convergence of the Newton iterations for the original discretization. In more difficult cases where it may be essential to have a good initial guess for the solution the quotient of the amount of work for the unperturbed problem to the amount of work for the perturbed problem may be much bigger.

No experiments have been done with iterative improvement for this class of problems.

4.3 Two-dimensional elliptic boundary value problems

We will discuss two classes of problems on rectangular regions. All problems discussed are assumed to have unique smooth solutions, however some numerical results for a problem with a non-smooth solution will be presented.

The existence of smooth error expansions for discretizations of elliptic problems are discussed in Volkov (1957), Stetter (1965), Hofman (1967) and Pereyra (1970).

4.3.1 Problems non-linear in u only

Consider

$$\nabla^2 u = f(x, y, u) \qquad (x, y) \in \Omega$$

$$u = h(x, y) \qquad (x, y) \in \partial\Omega$$

where

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

$$\Omega = \{0 \leq x \leq 1,\ 0 \leq y \leq 1\}$$

$\partial\Omega$ the boundary of Ω

Assume that f and h are such that the solution $u \in C^{2s}(\Omega)$.

In our operator formalism we have

$$F: \ E \to E^0$$

$$E = C^{2s}(\Omega) \quad ; \quad E^0 = C(\Omega) \times C(\partial\Omega)$$

with the norms

$$||z||_E = \max_{(x, y) \in \Omega} \sum_{v=0}^{2s} \frac{1}{v!} \sum_{\mu=0}^{v} \left|\frac{\partial^v z(x, y)}{\partial x^\mu \partial y^{v-\mu}}\right|$$

$$||g||_{E^0} = \max_{(x, y) \in \Omega} |g^1(x, y)| + \max_{(x, y) \in \partial\Omega} |g^0(x, y)|$$

where

$$g = \begin{pmatrix} g^1(x, y) & (x, y) \in \Omega \\ g^0(x, y) & (x, y) \in \partial\Omega \end{pmatrix}$$

$$F(z) = \begin{pmatrix} \nabla^2 z - f(x, y, z) & (x, y) \in \Omega \\ z - h(x, y) & (x, y) \in \partial\Omega \end{pmatrix}$$

Introduce $\Delta_h: E \to E_h \qquad E_h = R^{(N+1)^2}$

$$\Delta_h z = (z(x_i, y_j) \qquad i = 0(1)N,\ j = 0(1)N)$$

$$x_i = i \cdot h \qquad i = 0(1)N$$

$$y_j = j \cdot h \qquad j = 0(1)N$$

$$h = 1/N$$

$$||\xi||_{E_h} = \max_{\substack{0 \le i \le N \\ 0 \le j \le N}} |\xi_{ij}|$$

and

$$\Delta_h^0: E^0 \to E_h^0 \quad ; \quad E_h^0 = R^{(N-1)^2} \times R^{N+1} \times R^{N+1} \times R^N \times R^N$$

$$\Delta_h^0 g = \begin{bmatrix} g^1(x_i, y_j) & \{^{i = 1(1)N-1}_{j = 1(1)N-1} \\ g^0(x_0, y_j) & j = 0(1)N \\ g^0(x_N, y_j) & j = 0(1)N \\ g^0(x_i, y_0) & i = 1(1)N-1 \\ g^0(x_i, y_N) & i = 1(1)N-1 \end{bmatrix}$$

for $g \in E^0$ as above.

$$||\gamma||_{E_h^0} = \max[|\gamma_{ij}^1|;i = 1(1)N-1,\ j = 1(1)N-1,$$

$$|\gamma_j^{01}|;j = 0(1)N,\ |\gamma_j^{02}|;j = 0(1)N,$$

$$|\gamma_i^{03}|;i = 1(1)N-1,\ |\gamma_i^{04}|;i = 1(1)N-1]$$

where

$$\gamma = \begin{bmatrix} \gamma_{ij}^{1} & i = 1(1)N-1,\ j = 1(1)N-1 \\ \gamma_{j}^{01} & j = 0(1)N \\ \gamma_{j}^{02} & j = 1(1)N \\ \gamma_{i}^{03} & i = 1(1)N-1 \\ \gamma_{i}^{04} & i = 1(1)N-1 \end{bmatrix} \in E_h^0$$

Define

$$\phi_h: \ E_h \to E_h^0$$

$$\phi_h(\xi) = \begin{bmatrix} \dfrac{\xi_{i+1j} - \xi_{ij+1} - 4\xi_{ij} + \xi_{i-1j} + \xi_{ij-1}}{h^2} - f(x_i, y_j, \xi_{ij}) & \\ & i = 1(1)N-1,\ j = 1(1)N-1 \\ \xi_{0j} - h(x_0, y_j) & j = 0(1)N \\ \xi_{Nj} - h(x_N, y_j) & j = 0(1)N \\ \xi_{i0} - h(x_i, y_0) & i = 1(1)N-1 \\ \xi_{iN} - h(x_i, y_N) & i = 1(1)N-1 \end{bmatrix}$$

One can easily verify that

$$\phi_h(\Delta_h z) = \Delta_h^0\{F(z) + \sum_{j=2}^{2s-2} f_j(z)\, h^j\} + O(h^{2s-1})$$

Let

$$\phi_h^E(\xi) = \begin{bmatrix} \psi_{ij}(\xi) & i = 1(1)N-1,\ j = 1(1)N-1 \\ \xi_{0j} - h(x_0, y_j) & j = 0(1)N \\ \xi_{Nj} - h(x_N, y_j) & j = 0(1)N \\ \xi_{i0} - h(x_i, y_0) & i = 1(1)N-1 \\ \xi_{iN} - h(x_i, y_N) & i = 1(1)N-1 \end{bmatrix}$$

where $\psi_{ij}: E_h \to R$.

We will consider three different alternatives for ψ_{ij}:

1. $$\psi_{ij}(\xi) = DX^{2,4}_{x_i,y_j}(\xi) + DY^{2,4}_{x_i,y_j}(\xi) - f(x_i, y_j, \xi_{ij})$$

where DX and DY are operators of the kind discussed in section 3. It is easy to verify that

$$\phi^E_h(\Delta_h z) = \Delta^0_h\{F(z)\} + O(h^4)$$

for all $z \in E$.

2. $$\psi_{ij}(\xi) = (\xi_{i-1\ j-1} + 4\xi_{i-1j} + \xi_{i-1\ j+1} + 4\xi_{ij-1} - 20\xi_{ij} + 4\xi_{ij+1}$$
$$+ \xi_{i+1\ j-1} + 4\xi_{i+1j} + \xi_{i+1\ j+1})/(6h^2)$$
$$- \frac{1}{72}[f(x_{i-1}, y_{j-1}, \xi_{i-1,j-1}) + 4f(x_{i-1}, y_j, \xi_{i-1,j})$$
$$+ f(x_{i-1}, y_{j+1}, \xi_{i-1,j+1}) + 4f(x_i, y_{j-1}, \xi_{ij-1})$$
$$- 20f(x_i, y_j, \xi_{ij}) + 4f(x_i, y_{j+1}, \xi_{i,j+1})$$
$$+ f(x_{i+1}, y_{j-1}, \xi_{i+1,j-1}) + 4f(x_{i+1}, y_j, \xi_{i+1,j})$$
$$+ f(x_{i+1}, y_{j+1}, \xi_{i+1,j+1})] - f(x_i, y_j, \xi_{ij})$$

One can show that for the solution u of $F(u) = 0$

$$\psi_{ij}(\Delta_h u) = \nabla^2 u(x_i, y_j) - f(x_i, y_j, u(x_i, y_j)) + O(h^4)$$

so

$$\phi^E_h(\Delta_h u) = \Delta^0_h\{F(u)\} + O(h^4)$$

for the solution u of $F(u) = 0$. In this case we have to rely on theorem 2, while for perturbations 1 and 3 we can rely on both theorem 1 and theorem 2.

Further one can show that

$$\phi^E_h(\Delta_h u) = \Delta^0_h\{F(u) + \sum_{j=4}^{2s-2} f_{E,j}(u)\cdot h^j\} + O(h^{2s-1})$$

We have used the ninepoint operator ∇_9^2 to approximate the Laplace operator ∇^2. For notation and the results below see e.g. p. 321 in Bjorck, Dahlquist (1973).

We have

$$\nabla_9^2 u = \nabla^2 u + \frac{h^2 \nabla^4 u}{12} + O(h^4)$$

But $\nabla^2 u = f(x, y, u)$ so

$$\nabla^2 u = f(x, y, u) + h^2 \nabla^2 f(x, y, u)/12 + O(h^4)$$

Thus if we use the ninepoint operator to approximate $\nabla^2 f(x, y, u)$ we get

$$\nabla_9^2 u - f(x, y, u) - \frac{h^2}{12} \nabla_9^2 f(x, y, u) = O(h^4)$$

which gives the result above.

3. $$\psi_{ij}(\xi) = D^{2,q}_{x_i y_j}(\xi) + DY^{2,q}_{x_i y_j}(\xi) - f(x_i, y_j, \xi_{ij})$$

then

$$\phi_h^E(\Delta_h z) = \Delta_h^0\{F(z)\} + O(h^q) .$$

so if $q \geq 2\cdot(v_{max} + 1)$ we can take

$$\phi_{h,v} \equiv \phi_h^E \quad ; \quad v = 1, 2, \ldots, v_{max}$$

and make v_{max} iterative improvements.

Numerical results

The problem above with $f = 0$ and $h(x, y) = \sin(x) \sin h(y) + \cos h(x) \cos y - (x^2 - y^2)/2$ was discretized as above (Kronsjo, Dahlquist (1972)). The system of linear equations obtained was solved iteratively with SOR with optimal relaxation parameter, and to get an initial approximation we interpolated the boundary values linearly. The iterations were terminated when the residual was less than ε.

The perturbation operators of both 1 and 2 above were used. The perturbed problem was solved with the same method but now we used the solution from the unperturbed problem as initial approximation and the iterations were terminated if an estimate of the iteration error was either less than γ times the estimated discretization error or less than ε. The errors were measured in the maximum norm.

The calculations were performed in double precision on an IBM 360/75 with N = 10 and some different values of ε and γ. The entries in the column "Number of iterations" stand for the number of iterations for the unperturbed problem/the number of iterations for the perturbed problem.

Perturbation	ε	γ	Actual Error	Estimated Error	Number of Iterations
1	10^{-7}	0.1	$1.51 \cdot 10^{-4}$	$1.40 \cdot 10^{-4}$	32/7
	10^{-10}	10^{-4}	$1.51 \cdot 10^{-4}$	$1.50 \cdot 10^{-4}$	44/21
2	10^{-4}	10^{-4}	$1.27 \cdot 10^{-4}$	$5.3 \cdot 10^{-5}$	21/1
	10^{-5}	10^{-4}	$1.47 \cdot 10^{-4}$	$1.36 \cdot 10^{-4}$	24/8
	10^{-6}	10^{-4}	$1.51 \cdot 10^{-4}$	$1.50 \cdot 10^{-4}$	29/14
	10^{-7}	10^{-4}	$1.51 \cdot 10^{-4}$	$1.50 \cdot 10^{-4}$	32/18
	10^{-7}	0.1	$1.51 \cdot 10^{-4}$	$1.40 \cdot 10^{-4}$	32/7
	10^{-10}	10^{-4}	$1.51 \cdot 10^{-4}$	$1.50 \cdot 10^{-4}$	44/21

Note that there is no difference between the results for the two different perturbations. Further note that to get a reliable error estimate we must get the iteration error sufficiently small. The estimation algorithm can only estimate the discretization error, not the iteration error. In fact, there is a danger of amplification of the iteration errors when we apply

the operator ϕ_h^E to the solution of the unperturbed problem. Note also that we can save some work by terminating the iterations for the perturbed problem early, i.e. by choosing a larger value for γ.

Analogous results were obtained for Laplace equation with derivative boundary conditions on some of the sides of the unit square.

The same problem was solved with iterative improvement, using the perturbation operators $\phi_{h,v}$, v = 1, 2, ... according to point 3 above. We used $\varepsilon = 10^{-14}$, $\gamma = 10^{-9}$ and some different values of N and q. The tables below give the maximal errors in the successive iterates for the iterative improvement.

N = 10 q = 8	Iteration number						
	0	1	2	3	4	5	6
Error	$1.5\ 10^{-4}$	$1.2\ 10^{-6}$	$5.4\ 10^{-8}$	$7.8\ 10^{-9}$	$1.4\ 10^{-9}$	$2.7\ 10^{-10}$	$5.6\ 10^{-11}$
Number of Iterations in SOR	52	40	37	31	26	24	22

N = 10 q = 4	Iteration number						
	0	1	2	3	4	5	6
Error	$1.5\ 10^{-4}$	$1.2\ 10^{-6}$	$6.6\ 10^{-8}$	$6.9\ 10^{-8}$	$6.9\ 10^{-8}$	$6.9\ 10^{-8}$	$6.9\ 10^{-8}$
Number of Iterations in SOR	52	40	37	30	24	20	19

N = 20 q = 12	Iteration number						
	0	1	2	3	4	5	6
Error	$3.8\ 10^{-5}$	$7.6\ 10^{-8}$	$4.2\ 10^{-9}$	$9.0\ 10^{-10}$	$2.7\ 10^{-10}$	$9.9\ 10^{-11}$	$4.0\ 10^{-11}$
Number of Iterations in SOR	102	80	62	51	48	45	43

N = 20 q = 8	Iteration number						
	0	1	2	3	4	5	6
Error	$3.8\ 10^{-5}$	$7.6\ 10^{-8}$	$3.4\ 10^{-9}$	$5.2\ 10^{-10}$	$9.8\ 10^{-11}$	$2.1\ 10^{-11}$	$4.4\ 10^{-12}$
Number of Iterations in SOR	102	80	62	51	47	43	41

Note 1 In all cases but one (q = 12) we have made more iterative improvements than suggested by theorems 3 and 4. In the theorems it is implied that q should be chosen such that $q \geq$ (maximal number of iterations + 1)·2. Further note that we improve the accuracy of our solutions even after the suggested maximal number of iterations has been exceeded. For q = 4 e.g., only one improvement is reasonable according to the theorems, but the second iterate has a considerably much smaller error than the first. This astonishing result is a lucky coincidence in the construction of the perturbation operators $\phi_{h,v}$, v = 1, 2, ... For almost all the gridpoints (all but the points closest to the boundary points) the formulas we use are of order 6 (rather than of order 4) for all functions u such that $\nabla^2 u = 0$. A similar result holds for q = 8. If $\phi_{h,v}(\Delta_h u) = O(h^q)$ we cannot increase the order of accuracy of our solutions by making extra iterations, however the "error constant" can in some cases be decreased in this way.

Note 2 In the cases q = 8 and q = 12 the maximal errors in the last three iterations occur close to the boundary. At the interior points the errors are much smaller. The explanation for this is that for the points close to the boundary we have to use non-centered approximations to the derivatives, while at the interior points we can use symmetric approximations.

Note 3 There is a certain amplification of the iteration errors when $\phi_{h,v}$ is applied to the solution. The larger q is the larger is this

amplification. Further the larger q is the larger are the "error constants" for the formulas employed close to the boundaries. These facts may explain why we get better results with q = 8 than with q = 12 for N = 20.

4.3.2 The minimal surface equation

As an exercise in how to proceed in a more difficult case we consider

$$\frac{\partial}{\partial x}\left[\frac{1}{\sqrt{1 + u_x^2 + u_y^2}}\frac{\partial u}{\partial x}\right] + \frac{\partial}{\partial y}\left[\frac{1}{\sqrt{1 + u_x^2 + u_y^2}}\frac{\partial u}{\partial y}\right] = 0 \qquad (x, y) \in \Omega$$

$$u = (\cos h^2(y) - x^2)^{1/2} \qquad (x, y) \in \partial\Omega$$

where

$$\Omega = \{0 \leq x \leq 1, 0 \leq y \leq 1\}$$

In vector operator notation the equation reads

$$\nabla\cdot[\gamma(|\nabla u|^2)\nabla u] = 0$$

where

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$$

$$\gamma(|\nabla u|^2) = (1 + |\nabla u|^2)^{-1/2}$$

Note that the exact solution is

$$u(x, y) = (\cos h^2(y) - x^2)^{1/2}$$

and that u_x has a singularity at x = 1, y = 0. The problem is discretized and solved according to Concus (1967) in the following way:

Introduce the gridpoints

$$x_i = i\cdot h \qquad i = 0(1)N$$

$$y_j = j\cdot h \qquad j = 0(1)N$$

where h = 1/N.

Approximate the differential equation by

$$f_{ij} = \gamma_{\overline{ij}}(2u_{ij} - u_{i-1,j} - u_{i,j-1}) + \gamma_{\overline{i+1,j}}(2u_{ij} - u_{i+1,j} - u_{i,j-1})$$
$$+ \gamma_{\overline{i,j+1}}(2u_{ij} - u_{i-1,j} - u_{i,j+1}) + \gamma_{\overline{i+1,j+1}}(2u_{ij} - u_{i+1,j} - u_{i,j+1})$$
$$= 0 \quad , \quad i = 1(1)N - 1, \; j = 1(1)N - 1$$

where $\gamma_{\overline{ij}} = \gamma(|\nabla u|^2_{ij})$ denotes γ for the mesh cell with center $(i - 1/2, j - 1/2)$, which is evaluated by use of

$$|\nabla u|^2_{ij} = \frac{1}{2h^2}[(u_{ij} - u_{i-1,j})^2 + (u_{ij} - u_{i,j-1})^2$$
$$+ (u_{i,j-1} - u_{i-1,j-1})^2 + (u_{i-1,j} - u_{i-1,j-1})^2].$$

This approximation is consistent of order 2 with the differential equation. The boundary conditions are represented by

$$u_{ij} = [\cos h^2(y_j) - x_i^2)^{1/2} \qquad (x_i, y_j) \in \partial\Omega$$

The system of nonlinear equations above are solved iteratively by computing u_{ij}^{k+1}, the $(k + 1)$th approximation to u_{ij} from

$$u_{ij}^{k+1} = u_{ij}^{k} - \omega \frac{f_{ij}[u_{11}^{k+1}, \ldots, u_{i-1,j}^{k+1}, u_{ij}^{k}, \ldots, u_{N,N-1}^{k}]}{\frac{\partial f_{ij}}{\partial u_{ij}}[u_{11}^{k+1}, \ldots, u_{i-1,j}^{k+1}, u_{ij}^{k}, \ldots, u_{N,N-1}^{k}]},$$

where ω is the relaxation parameter. The initial approximation was obtained by linear interpolation of the boundary values and the iterations were terminated when the last correction was less than ε.

The following basic discretization formula was used for the perturbation operator ϕ_h^E:

Introduce the notation

$$\xi = (\xi_{ij} \qquad i = 0(1)N, \; j = 0(1)N)$$

and

$$px(\xi) = \left(\frac{DX^{1,4}_{x_i,y_j}(\xi)}{\sqrt{1 + (DX^{1,4}_{x_i,y_j}(\xi))^2 + (DY^{1,4}_{x_i,y_j}(\xi))^2}} \qquad \begin{matrix} i = 0(1)N \\ j = 0(1)N \end{matrix} \right)$$

$$py(\xi) = \left(\frac{DY^{1,4}_{x_i,y_j}(\xi)}{\sqrt{1 + (DX^{1,4}_{x_i,y_j}(\xi))^2 + (DY^{1,4}_{x_i,y_j}(\xi))^2}} \qquad \begin{matrix} i = 0(1)N \\ j = 0(1)N \end{matrix} \right)$$

Here DX and DY are operators of the type described in section 3. Further

$$\psi_{ij}(\xi) = DX^{1,4}_{x_i,y_j}(px(\xi)) + DY^{1,4}_{x_i,y_j}(py(\xi)) \qquad i = 1(1)N - 1,\ j = 1(1)N - 1$$

Note that with proper definition of Δ_h,

$$\psi_{ij}(\Delta_h z) = [\frac{\partial}{\partial x}(\frac{1}{\sqrt{1 + u_x^2 + u_y^2}} \frac{\partial u}{\partial x}) + \frac{\partial}{\partial y}(\frac{1}{\sqrt{1 + u_x^2 + u_y^2}} \frac{\partial u}{\partial y})](x_i, y_j) + O(h^4)$$

ψ_{ij} is the basic discretization operator for the perturbations. The perturbed problem was solved with the same iterative method as the unperturbed problem, but now we used the solution of the unperturbed problem as initial approximation and terminated the iterations when an estimate of the iteration error was either less than γ times the estimated discretization error or less than ε. All quantities were measured in the maximum norm.

The problem was solved in double precision on an IBM 360/75 and we used $\varepsilon = 10^{-8}$, $\gamma = 10^{-4}$ and some different values of N and ω. The entries in the column, "Number of Iterations," refer to the number of iterations for unperturbed problem/number of iterations for perturbed problem.

N	ω	Actual Error	Estimated Error	Number of Iterations
10	1.7	$4.3 \cdot 10^{-3}$	$8.9 \cdot 10^{-3}$	55/28
40	1.87	$2.3 \cdot 10^{-3}$	$5.0 \cdot 10^{-3}$	147/71

The maximum error was obtained in the vicinity of $x = 1$, $y = 0$. Due to the singularity of u_x at that point we cannot expect to get very good results (or good error estimates) in that region. However, for the rest of the unit square much better results and error estimates were obtained. Below the results at some representive points are tabulated for $N = 10$ and $\omega = 1.7$.

x	y	Solution	Actual Error	Estimated Error
0.2	0.6	1.17	$1.10 \cdot 10^{-4}$	$1.16 \cdot 10^{-4}$
0.5	0.2	0.89	$1.13 \cdot 10^{-4}$	$1.90 \cdot 10^{-4}$
0.5	0.6	1.08	$4.64 \cdot 10^{-4}$	$4.50 \cdot 10^{-4}$
0.8	0.2	0.63	$1.24 \cdot 10^{-3}$	$1.71 \cdot 10^{-3}$
0.8	0.6	0.88	$1.48 \cdot 10^{-3}$	$1.38 \cdot 10^{-3}$
0.9	0.3	0.54	$4.1 \cdot 10^{-3}$	$8.9 \cdot 10^{-3}$

Analogous results were obtained for the same problem with $\partial u/\partial x = 0$ at $x = 0$ and unchanged boundary conditions on the remaining sides of the unit square.

4.4 Parabolic partial differential equations

Consider

$$u_t = f(t, x, u, u_x, u_{xx}) \qquad (t, x) \in \Omega$$

$$u(a, t) = f_1(t), \quad u(b, t) = f_2(t), \quad t > 0 \; ; \; u(x, 0) = h(x), \quad a \leq x \leq b$$

where $\Omega = \{a < x < b, \; t > 0\}$. Assume that f, f_1, f_2 and h are such that $u \in C^s(\Omega)$.

The definition of spaces, norms and operators for our operator formalism is left as an exercise for the reader.

We will use the method of lines to discretize the problem in space and then solve the system of ordinary differential equations so obtained with two simple methods; the explicit Euler method and the implicit backward Euler method. It is plausible that for sufficiently small h (discretization parameter) there exists a smooth expansion of the global discretization error for both these methods (Stetter (1965), Keller (1970)). However, the main advantage of implicit methods is that one can use large time steps and still obtain accurate results. For realistic stepsizes in time there are, to my knowledge, no general results on the existence of smooth expansions for the global discretization error. The system of differential equations that was obtained by the method of lines is stiff and hence results on error expansions for numerical methods for stiff systems of ordinary differential equations may be useful for our methods. For such problems a discussion of error expansions for large values of the stepsize h (not only asymptotically) for implicit one-step methods can be found in Dahlquist, Lindberg (1973). These questions warrant further study that are outside the scope of this report.

Here we assume the existence of sufficiently smooth error expansions for the methods and stepsizes we use and proceed under that assumption to estimate the global discretization error with our algorithm. Numerical results indicate that the assumption is reasonable.

4.4.1 The method of lines with Euler's method

Discretize the problem according to

$$x_i = a + i \cdot h; \quad i = 0, 1, \ldots, N; \quad h = \frac{b - a}{N}$$

$$t_j = j \cdot k; \quad j = 0, 1, \ldots; \quad k = c \cdot h^2 \quad c \leq 0.5$$

$$u_{ij} \approx u(x_i, t_j)$$

$$\frac{u_{ij+1} - u_{ij}}{k} - f(t_j, x_i, u_{ij}, \frac{u_{i+1j} - u_{i-1j}}{2h}, \frac{u_{i+1j} - 2u_{ij} + u_{i-1j}}{2h}) = 0$$

$$i = 1(1)N - 1, \; j = 0, 1, \ldots$$

$$u_{i0} - h(x_i) = 0 \quad i = 0(1)N$$

$$u_{0j} - f_1(t_j) = 0 \quad u_{Nj} - f_2(t_j) = 0 \quad j = 1, 2, \ldots$$

This is an explicit method so we simply proceed one time-level a time computing according to the formula above.

The perturbations are defined according to:

Introduce $\xi = (\xi_{ij} \quad i = 0(1)N, \; j = 0, 1, \ldots)$

$$\psi_{ij}(\xi) = DT^{1,2}_{x_i,t_j}(\xi) - f(t_j, x_i, \xi_{ij}, DX^{1,4}_{x_i,t_j}(\xi), DX^{2,4}_{x_i,t_j}(\xi))$$

$$i = 1(1)N - 1, \; j = 0, 1, \ldots$$

Here the operators $DT^{v,m}_{x,t}$, $DX^{v,m}_{x,t}$ are of the type discussed in section 3.

For the interior grid points we have e.g.

$$DT^{1,2}_{x_i,t_j}(\xi) = \frac{\xi_{ij+1} - \xi_{ij-1}}{2k}$$

$$DX^{1,4}_{x_i,t_j}(\xi) = \frac{\xi_{i-2j} - 8\xi_{i-1j} + 8\xi_{i+1j} - \xi_{i+2j}}{12h}$$

The perturbed problem is solved in parallel with the unperturbed problem.

Several alternative perturbations exist, cf. section 4.1.

The problem below was solved on an IBM 360/75 using double precision arithmetic. We used $N = 10$, and $c = 2.5/\pi^2$.

$$u_t = u_{xx} \qquad 0 < x < \pi,\ t > 0 \qquad u(0, t) = u(\pi, t) = 0$$

with the following initial value functions h(x)

(1) $$h(x) = \sin(x)$$

exact solution: $u(x, t) = e^{-t} \cdot \sin(x)$

(2) $$h(x) = x(\pi - x)$$

exact solution: $$u(x, t) = \frac{8}{\pi} \sum_{n=1}^{\infty} (2n - 1)^{-3}\, e^{-(2n-1)^2 t} \sin(2n - 1)x$$

(3) $$h(x) = \begin{cases} x & 0 \leq x \leq \pi/2 \\ \pi - x & \pi/2 < x \leq \pi \end{cases}$$

exact solution: $$u(x, t) = \frac{4}{\pi} \sum_{m=1(2)}^{\infty} (-1)^{(m-1)/2} \cdot \frac{1}{m^2} \sin(m\, x) \cdot e^{-m^2 t}$$

The tables below give the maximal errors for selected time levels for the different problems.

Problem #1 N = 10, k = 0.025

t	Actual Error	Estimated Error
0.05	$2.05 \cdot 10^{-4}$	$2.13 \cdot 10^{-4}$
0.20	$7.05 \cdot 10^{-4}$	$7.46 \cdot 10^{-4}$
0.50	$1.30 \cdot 10^{-3}$	$1.35 \cdot 10^{-3}$
0.75	$1.52 \cdot 10^{-3}$	$1.56 \cdot 10^{-3}$

Problem #2 N = 10, k = 0.025

t	Actual Error	Estimated Error
0.05	$2.6 \cdot 10^{-3}$	$8.4 \cdot 10^{-3}$
0.20	$2.2 \cdot 10^{-3}$	$3.4 \cdot 10^{-3}$
0.50	$3.2 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$
0.75	$3.9 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$

Problem #3 N = 10, k = 0.025

t	Actual Error	Estimated Error
0.05	$3.7 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$
0.20	$1.6 \cdot 10^{-2}$	$4.9 \cdot 10^{-3}$
0.50	$9.8 \cdot 10^{-3}$	$3.1 \cdot 10^{-3}$
0.75	$8.3 \cdot 10^{-3}$	$3.1 \cdot 10^{-3}$

For these problems the amount of work needed to get the error estimate is approximately equal to the amount of work needed to solve the unperturbed problem.

From the numerical results it is obvious that the estimates are not very reliable for test problem 3 where the initial value function is not smooth; for the other two problems the error estimates are quite good.

4.4.2 The method of lines with the backward Euler method

Discretize the problem according to

$$x_i = a + ih; \quad i = 0(1)N \quad h = (b - a)/N$$

$$t_j = j \cdot k; \quad j = 0, 1, \ldots; \quad k = c \cdot h$$

$$u_{ij} \approx u(x_i, t_j)$$

$$\frac{u_{ij+1} - u_{ij}}{k} - f\left(t_{j+1}, x_i, u_{ij+1}, \frac{u_{i+1\,j+1} - u_{i-1\,j+1}}{2h}, \right.$$

$$\left. \frac{u_{i+1\,j+1} - 2u_{ij+1} + u_{i-1\,j+1}}{h^2}\right) = 0$$

$$i = 1(1)N - 1, \; j = 1, 2, \ldots$$

$$u_{i0} - h(x_i) = 0 \qquad i = 0(1)N$$

$$u_{0j} - f_1(t_j) = 0 \quad u_{Nj} - f_2(t_j) = 0 \qquad j = 1, 2, \ldots$$

The system of nonlinear equations that is obtained in each time step is solved by Newton iteration. As initial guess we use the solution at the previous time level and the iterations are terminated when the last correction is less than ε. The maximum norm is used to measure all quantities.

Introduce the notation

$$\xi = (\xi_{ij}; \quad i = 0(1)N, \; j = 0, 1, \ldots)$$

The basic discretization for the perturbation operator is

$$\psi_{ij}(\xi) = \frac{\xi_{ij+1} - \xi_{ij-1}}{2k} - f(t_j, x_i, \xi_{ij}, DX^{1,4}_{x_i,t_j}(\xi), DX^{2,4}_{x_i,t_j}(\xi))$$

$$i = 1(1)N - 1 \qquad j = 1, 2, \ldots$$

The operators $DX^{v,m}_{x,t}$ are of the type discussed in section 3.

For the perturbed problem we use modified Newton iteration to solve the system of nonlinear equations obtained in each time step. As initial approximation we use the solution of the unperturbed problem and terminate the iterations when an estimate of the iteration error is either less than γ times the estimated discretization error at the current time level, or less than ε.

The problems below were solved in this fashion.

(1) $$u_t = (2 + \sin(u_{xx}))u_{xx} + (1 - \sin(u))u$$

$$0 < x < \pi/2, \; t > 0$$

$$u(0,t) = 0 \quad ; \quad u(\pi/2, t) = e^{-t} \qquad t > 0$$

$$u(x, 0) = \sin(x) \qquad 0 \leq x \leq \pi/2$$

exact solution: $u(x, t) = e^{-t} \sin(x)$

(Kolar (1972))

(2, 3, 4) Burger's equation

$$u_t = - u\, u_x + v\, u_{xx} \qquad 0 < x < \pi,\ t > 0$$

$$u(0, t) = u(\pi, t) = 0$$

$$u(x, 0) = \sin(x)$$

exact solution:

$$u(x, t) = 4v \frac{\sum_{n=1}^{\infty} \exp(- v\, n^2\, t) n\, I_n(\frac{1}{2v}) \cdot \sin(n\, x)}{I_0(\frac{1}{2v}) + 2 \sum_{n=1}^{\infty} \exp(- v\, n^2\, t) I_n(\frac{1}{2v}) \cdot \cos(n\, x)}$$

The functions I_n, $n = 0, 1, \ldots$ are the modified Bessel functions. (Cole (1951)). We have used the following values of v:

(2) $v = 1$

(3) $v = 1/8$

(4) $v = 1/16$

In the calculations for the tables below we have used $\varepsilon = 10^{-6}$, $k = 0.1$ and some different values of N. The computations were done in double precision on an IBM 360/75. The maximum errors at some different time levels and the number of iterations for the unperturbed/perturbed problems are recorded in the tables.

Problem #1, N = 10, k = 0.1

t	Actual Error	Estimated Error	Number of iterations
0.1	$3.2 \cdot 10^{-3}$	$2.6 \cdot 10^{-3}$	5/2
0.5	$6.6 \cdot 10^{-3}$	$5.8 \cdot 10^{-3}$	6/2
1.0	$4.0 \cdot 10^{-3}$	$4.2 \cdot 10^{-3}$	6/2
1.5	$2.22 \cdot 10^{-3}$	$2.19 \cdot 10^{-3}$	5/2
2.0	$1.22 \cdot 10^{-3}$	$1.21 \cdot 10^{-3}$	5/2

Problem #2, N = 20, k = 0.1

t	Actual Error	Estimated Error	Number of Iterations
0.1	$1.1 \cdot 10^{-2}$	$7.5 \cdot 10^{-3}$	3/2
0.5	$1.9 \cdot 10^{-2}$	$1.6 \cdot 10^{-2}$	3/2
1.0	$2.0 \cdot 10^{-2}$	$1.8 \cdot 10^{-2}$	3/2
2.0	$1.4 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$	3/2
4.0	$3.9 \cdot 10^{-3}$	$3.8 \cdot 10^{-3}$	2/2
6.0	$8.3 \cdot 10^{-4}$	$8.5 \cdot 10^{-4}$	2/2
8.0	$1.6 \cdot 10^{-4}$	$1.7 \cdot 10^{-4}$	2/2

Problem #3, N = 20, k = 0.1

t	Actual Error	Estimated Error	Number of Iterations
0.1	$4.3 \cdot 10^{-3}$	$3.8 \cdot 10^{-3}$	3/2
0.5	$1.5 \cdot 10^{-2}$	$1.4 \cdot 10^{-2}$	3/2
1.0	$1.8 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	3/2
2.0	$1.7 \cdot 10^{-2}$	$1.6 \cdot 10^{-2}$	3/2
4.0	$2.5 \cdot 10^{-2}$	$2.6 \cdot 10^{-2}$	3/2
6.0	$1.5 \cdot 10^{-2}$	$1.6 \cdot 10^{-2}$	3/2
8.0	$9.7 \cdot 10^{-3}$	$9.9 \cdot 10^{-3}$	3/2

Problem #4, N = 20, k = 0.1

t	Actual Error	Estimated Error	Number of Iterations
0.1	$4.5 \cdot 10^{-3}$	$4.1 \cdot 10^{-3}$	3/2
0.5	$1.8 \cdot 10^{-2}$	$1.7 \cdot 10^{-2}$	3/2
1.0	$2.6 \cdot 10^{-2}$	$2.3 \cdot 10^{-2}$	3/2
2.0	$1.9 \cdot 10^{-1}$	$3.5 \cdot 10^{-1}$	3/2
4.0	$1.1 \cdot 10^{-1}$	$1.4 \cdot 10^{-1}$	3/2
6.0	$3.9 \cdot 10^{-2}$	$3.8 \cdot 10^{-2}$	3/2
8.0	$2.1 \cdot 10^{-2}$	$2.2 \cdot 10^{-2}$	3/2

All the estimates above are very good, except for problem #4 where for $t \approx 2.0$ we get estimates that are approximately twice the actual error. This problem developes a fairly sharp front with large derivatives with respect to x. When the front disappears the estimates become quite reliable again.

Note that the amount of work needed to find the error estimate is quite small compared to the amount of work needed to find the approximate solution.

For large values of N we need approximately $N^3/3^*$ (number of iterations for the unperturbed problem) operations to find the approximate solution and N^{2*} (number of iterations for the perturbed problem) operations to get the error estimate.

4.5 Hyperbolic partial differential equations

For the commonly used discretizations of initial-boundary value problems for hyperbolic partial differential equations, I know of no extensive discussion of the existence of smooth expansions of the global

discretization error. Some very interesting results are given in Gourlay, Morris (1968) and Skollermo (1975a, 1975b). In essence the main difficulty seems to be how to represent the numerical boundary conditions (i.e. the extra boundary conditions imposed by the numerical scheme) so the errors due to that representation (these errors are not smooth) is sufficiently small.

For the method of characteristics some authors have used extrapolation to increase the accuracy, see Lister (1960), Werner (1968), Smith (1970).

The results of this section are of an experimental nature, and no rigorous mathematical analysis is attempted. The numerical results indicate that by careful choice of the perturbation operators one can obtain good error estimates for problems with smooth solutions. A further study of this class of problems may prove very fruitful.

Consider the initial-boundary value problem

$$u_t = a\, u_x \qquad t \geq 0,\ 0 \leq x \leq 1$$

$$a > 0$$

$$u(x, 0) = f(x) \qquad 0 \leq x \leq 1$$

$$u(1, t) = h(t) \qquad t \geq 0$$

and use the discretization

$$x_i = i \cdot h \qquad i = 0, 1, \ldots, N \quad ; \quad h = 1/N$$

$$t_j = j \cdot k \qquad k = 0, 1, \ldots \quad ; \quad k = c \cdot h$$

$$u_{ij} \approx u(x_i, t_j)$$

$$\frac{u_{ij+1} - u_{ij-1}}{2k} - a\,\frac{u_{i+1j} - u_{i-1j}}{2h} = 0$$

$$u_{i0} = f(x_i) \qquad i = O(1)N$$

$$u_{Nj} = h(t_j) \qquad j = 0, 1, \ldots$$

This is the leap-frog method. Note that to be able to compute the approximate solution with the formula above we must, in some way, find

$$u_{i1} \qquad i = 0(1)N \qquad \text{(extra starting values)}$$

and

$$u_{0j} \qquad j = 1, 2, \ldots \qquad \text{(numerical boundary values)}$$

If the extra starting values are computed with a sufficiently accurate method we have

$$u_{ij} = u(x_i, t_j) + h^2 e(x_i, t_j) + h^{\mu} \varepsilon_{ij} + R_{ij}$$

where

$$|R_{ij}| = O(h^m) \qquad m = \min(4, \mu + 1)$$

μ depends on the method we use to find the numerical boundary values

$\varepsilon_{ij} = (-1)^i C(x_i, t_j)$ where C is a smooth function

(adapted from Skollermo (1975a)).

We have made some numerical experiments with our estimation algorithm for this discretization.

To construct the perturbation operator ϕ_h^E we proceed in the following way:

Define

$$\xi = (\xi_{ij}, \; i = 0(1)N, \; j = 0, 1, \ldots)$$

$$\varepsilon = (\varepsilon_{ij}, \; i = 0(1)N, \; j = 0, 1, \ldots)$$

$$\delta(\xi) = ((\xi_{i+1j} - \xi_{i-1j})/2h, \; i = 1(1)N - 1, \; j = 0, 1, \ldots)$$

Note that when we apply δ to the non-smooth error term $h^{\mu} \cdot \varepsilon$ of the expansion above we get

$$\delta_{ij}(h^\mu \varepsilon) = (-1)^{i-1} \cdot C_x(x_i, t_j) \cdot h^\mu + O(h^{2+\mu})$$

As the basic discretization for the perturbation operator ϕ_h^E we now use

$$\psi_{ij}(\xi) = DT^{1,4}_{x_i,t_j}(\xi) - a\, D\tilde{X}_{x_i,t_j}(\delta(\xi))$$

where

$$D\tilde{X}_{x_i,t_j}(\delta(\xi)) = \sum_{s=\ell_i}^{u_i} \alpha_{is}\, \delta_{sj}(\xi)$$

is a linear combination of some of the components of $\delta(\xi)$ such that $D\tilde{X}_{x_i,t_j}(\delta(\Delta_h z))$ is consistent of order 4 with $z_x(x_i, t_j)$. The coefficients α_{is} are independent of h so

$$h^\mu \psi_{ij}(\varepsilon) = (-1)^{i-1}(-c_t(x_i, t_j) - a\, c_x(x_i, t_j))h^\mu + O(h^{\mu+4})$$

$$= O(h^\mu)$$

This choice of the perturbation operator ϕ_h^E insures that no serious loss of accuracy in the error estimate is caused by the irregular error term $h^\mu \cdot \varepsilon$.

The theorems of section 2 do not cover this kind of error expansions and perturbation operators, but they could easily be modified to do so.

The following problem

$$u_t = u_x \qquad 0 < x < 1,\ t > 0$$

$$u(x, 0) = \sin(2\pi x) \qquad 0 < x < 1$$

$$u(1, t) = \sin(2\pi t) \qquad t > 0$$

with the exact solution $u(x, t) = \sin(2\pi(x + t))$ was solved with the leap-frog method.

First we extended the solution to the half plane $t \geq 0$ and used the periodicity of the exact solution to find the numerical boundary values, i.e.

$$u_{0j} = u_{Nj} \qquad j = 1, 2, \ldots$$

We also used the periodicity to simplify the formulas for the perturbation operator ϕ_h^E.

This problem was included to see if our algorithm would work in the case where the numerical boundary condition did not introduce a non-smooth error term. Two ways of finding the extra starting values u_{i1}, $i = 0(1)N$ were tested,

I. $$u_{i1} = \sin(2\pi(x_i + k)) \qquad i = 0(1)N$$

i.e. values from the exact solution.

II. $$u_{i1} = u_{i0} + \frac{1}{2}\lambda(u_{i+1\,0} - u_{i-1\,0}) + \frac{1}{2}\lambda^2(u_{i+1\,0} - 2u_{i0} + u_{i-1\,0})$$

with $\lambda = k/h$

i.e. the Lax-Wendroff scheme.

In tables 1 and 2 the maximal errors for some time levels are given for these cases.

In the next experiment we did not use the periodicity of the exact solution, but used the following two formulas to find the numerical boundary values

A. $$u_{0j+1} = u_{0j} + \lambda(u_{1j} - u_{0j}) \quad ; \quad \lambda = k/h$$

i.e. the explicit Euler scheme.

B. $$u_{0j+1} + u_{1j+1} - \lambda(u_{1j+1} - u_{0j+1}) = u_{0j} + u_{1j} - \lambda(u_{1j} - u_{0j}) \qquad \lambda = k/h$$

i.e. the Box scheme.

The Lax-Wendroff scheme was used to get the extra starting values u_{i1}, $i = 0(1)N$. In tables 3 and 4 the maximal error for some time levels are recorded.

In all the calculations for the tables we used

$$h = 0.025 \qquad k = 0.01875$$

and the computations were carried out in double precision on an IBM 360/75.

Table 1, starting procedure I

t	Actual Error	Estimated Error
0.15	$1.38 \cdot 10^{-3}$	$1.41 \cdot 10^{-3}$
0.75	$8.31 \cdot 10^{-3}$	$8.27 \cdot 10^{-3}$
1.5	$1.70 \cdot 10^{-2}$	$1.71 \cdot 10^{-2}$
3.0	$3.41 \cdot 10^{-2}$	$3.42 \cdot 10^{-2}$

Table 2, starting procedure II

t	Actual Error	Estimated Error
0.15	$1.71 \cdot 10^{-3}$	$2.18 \cdot 10^{-3}$
0.75	$8.52 \cdot 10^{-3}$	$9.29 \cdot 10^{-3}$
1.5	$1.70 \cdot 10^{-2}$	$1.71 \cdot 10^{-2}$
3.0	$3.41 \cdot 10^{-2}$	$3.42 \cdot 10^{-2}$

Table 3, boundary scheme A

t	Actual Error	Estimated Error
0.15	$1.71 \cdot 10^{-3}$	$1.72 \cdot 10^{-3}$
0.75	$8.71 \cdot 10^{-3}$	$1.15 \cdot 10^{-2}$
1.5	$2.28 \cdot 10^{-2}$	$2.63 \cdot 10^{-2}$
3.0	$3.43 \cdot 10^{-2}$	$4.69 \cdot 10^{-2}$

Table 4, boundary scheme B

t	Actual Error	Estimated Error
0.15	$1.71 \cdot 10^{-3}$	$1.72 \cdot 10^{-3}$
0.75	$8.37 \cdot 10^{-3}$	$9.09 \cdot 10^{-3}$
1.5	$1.69 \cdot 10^{-2}$	$1.73 \cdot 10^{-2}$
3.0	$3.37 \cdot 10^{-2}$	$3.45 \cdot 10^{-2}$

The nonlinear problem

$$u_t = \frac{\partial}{\partial x}\left(\frac{1}{2} u^2\right) \qquad 0 < x < 1,\ t > 0$$

$$u(x, 0) = 1 - x \qquad 0 < x < 1$$

$$u(1, t) = 0 \qquad t > 0$$

with the exact solution

$$u(x, t) = (1 - x)/(1 + t)$$

(Gourlay, Morris (1968)) was solved with the leap-frog method. We used the Lax-Wendroff starting procedure and the Box scheme for calculation of the numerical boundary values.

The discretization error was estimated as above. For $h = 0.05$ and $k = 0.015$ the maximal errors for some time levels are recorded in the table below.

Table 5

t	Actual Error	Estimated Error
0.15	$3.1 \cdot 10^{-5}$	$3.3 \cdot 10^{-5}$
0.75	$3.1 \cdot 10^{-5}$	$3.6 \cdot 10^{-5}$
1.5	$2.2 \cdot 10^{-5}$	$3.2 \cdot 10^{-5}$
3.0	$1.9 \cdot 10^{-5}$	$3.2 \cdot 10^{-5}$
4.5	$2.0 \cdot 10^{-5}$	$3.9 \cdot 10^{-5}$
6.0	$1.9 \cdot 10^{-5}$	$4.6 \cdot 10^{-5}$
7.5	$2.0 \cdot 10^{-5}$	$6.0 \cdot 10^{-5}$

4.6 Integral Equations

First consider Fredholm's integral equations of the second kind

$$y(x) - \lambda \int_a^b K(x, t)\, y(t)\, dt - f(x) = 0$$

with the following discretization

$$t_i = x_i = a + i \cdot h, \; i = 0(1)N \quad ; \quad h = (b - a)/N$$

$$\phi_h(\xi) = (\xi_i - h\lambda \sum_{j=0}^{N} \alpha_j K(x_i, t_j)\, \xi_j - f(x_i) \quad ; \quad i = 0(1)N)$$

where $\alpha_0 = \alpha_N = 1/2$; $\alpha_j = 1$, $j = 1(1)N - 1$.

We have used the trapezoidal rule to approximate the integral in the equation above.

Construct the perturbation operators

$$\phi_{h,v}(\xi) = (\xi_i - h\lambda \sum_{j=0}^{N} \alpha_{jv} K(x_i, t_j)\, \xi_j - f(x_i) \quad ; \quad i = 0(1)N)$$

$$v = 1, 2, \ldots$$

Here the coefficients α_{jv} are such that

$$\sum_{j=0}^{N} \alpha_{jv}\, \psi(t_j) = \int_a^b \psi(t)\, dt + \sum_{j=2(v+1)}^{s} f_j(\psi)\, h^j + O(h^{s+1})$$

The system of linear equations obtained for the unperturbed problem is solved by Gaussian elimination. The LU factorization of the coefficient matrix is saved for use in the sequence of perturbed problems.

The unperturbed problems takes approximately $(N + 1)^3/3$ operations (ignoring the number of operations needed to compute $K(x_i, t_j)$, $i = 0(1)N$, $j = 0(1)N$ and $f(x_i)$, $i = 0(1)N$) while each of the iterations in the iterative improvement takes approximately $2(N + 1)^2$ operations (ignoring the number of operations needed to calculate the coefficients α_{jv} of the perturbation operators, approximately $3\ [2(v + 1)]^2$).

The α_{jv} are best computed as weights of a composite quadrature formula of order at least $2(v + 1)$. The length of the intervals for the basic quadrature formulas are $M \cdot h$, where M has to be a factor of N (M may be equal to N). The coefficients of the basic formulas can be obtained as the solution of Van der Monde systems of linear equations in the same way we obtained the coefficients of the differentiation formulas of section 3.

In Van der Sluis (1972) a discussion of asymptotic error expansions for quadrature formulas of this kind can be found, and in Laurent (1964), Stetter (1965) asymptotic expansions for the global discretization error for our method are discussed.

The problems below were solved on an IBM 360/75 in double precision. All problems are taken from Netravali, de Figueiredo (1974).

(1) $$y(x) - \int_0^1 xt\, y(t)\, dt - (e^x - 1) = 0$$

exact solution: $y(x) = e^x$

(2) $$y(x) - \int_0^1 xt\, y(t)\, dt - (\sin(\pi x) - x/\pi) = 0$$

exact solution: $y(x) = \sin(\pi x)$

(3) $$y(x) - \int_0^1 x^4 e^{xt}\, y(t)\, dt - (x - x^3[e^x(1 - \tfrac{1}{x}) + \tfrac{1}{x}]) = 0$$

exact solution: $y(x) = x$

(4) $$y(x) - \int_0^1 x^4 e^{xt}\, y(t)\, dt - (\sin(\pi x) - \frac{\pi x^4(e^x + 1)}{x^2 + \pi^2}) = 0$$

exact solution: $y(x) = \sin(\pi x)$

We have used N = 16 and record below the maximum errors in the successive iterates.

Problem	Error in Iterate Number						
	0	1	2	3	4	5	6
1	$2.2\ 10^{-3}$	$2.1\ 10^{-6}$	$2.1\ 10^{-9}$	$2.0\ 10^{-12}$	$1.1\ 10^{-14}$	$1.6\ 10^{-15}$	$7.1\ 10^{-15}$
2	$1.5\ 10^{-3}$	$1.4\ 10^{-6}$	$1.4\ 10^{-9}$	$2.7\ 10^{-11}$	$2.7\ 10^{-14}$	$4.8\ 10^{-15}$	$3.2\ 10^{-16}$
3	$2.3\ 10^{-3}$	$2.3\ 10^{-5}$	$2.0\ 10^{-7}$	$1.8\ 10^{-9}$	$1.6\ 10^{-11}$	$1.6\ 10^{-13}$	$7.5\ 10^{-15}$
4	$6.6\ 10^{-3}$	$5.7\ 10^{-5}$	$5.1\ 10^{-7}$	$4.8\ 10^{-9}$	$4.3\ 10^{-11}$	$3.7\ 10^{-13}$	$7.5\ 10^{-15}$

Note 1 The coefficients α_{jv}, $j = 0, 1, \ldots, N$, $v = 1, 2, \ldots, v_{max}$ can also be made independent of v if they are chosen as $\alpha_{jv} = w_j$, $j = 0, \ldots, N$, where

$$\sum_{j=0}^{N} w_j \psi(t_j) = \int_a^b \psi(t)\, dt + O(h^\mu)$$

and $\mu \geq (v_{max} + 1)\cdot 2$.

Note 2 The technique described above can be used for other discretizations of the integral equation, using e.g. non-uniform grid points x_i, $i = 0, 1, \ldots, N$. We can also use the same technique for other types of integral equations, e.g. nonlinear, as long as the assumptions of theorem 3 or 4 are satisfied. Essentially we require smoothness of the exact solution and a smooth error expansion for the solution of the discretized problem.

Consider nonlinear integral equations of the type

$$y(x) - \int_a^b K(x, t, y(x), y(t))\, dt - f(x) = 0$$

with the following discretization

$$t_i = x_i = a + i\cdot h \quad , \quad i = 0(1)N \quad , \quad h = (b - a)/N$$

$$\phi_h(\xi) = (\xi_i - h \sum_{j=0}^{N} \alpha_j K(x_i, t_j, \xi_i, \xi_j) - f(x_i) \quad ; \quad i = 0(1)N)$$

where

$$\alpha_0 = \alpha_N = 1/2 \quad ; \quad \alpha_j = 1 \quad , \quad j = 1(1)N - 1$$

The system of nonlinear equations obtained is solved by Newton iteration.

The perturbations are constructed analogously to the perturbations for the linear problems above, and the system of nonlinear equations obtained for each perturbed problem is solved by Newton iteration. The iterations are terminated when the iteration error becomes less than ε.

The problems below were solved in this fashion on an IBM 360/75 in double precision.

(1) $$y(x) - \int_0^1 x \cdot t \cdot y(t)^2 \, dt - 3x/4 = 0$$

exact solutions: $y(x) = x$ and $y(x) = 3x$

(Moore (1968))

(2) $$y(x) - \int_0^1 \frac{t^2}{1+x} y(x)^2 \, y(t)^2 \, dt - e^x - e^{2x}(e^2 - 1)/(4(1+x)) = 0$$

exact solution: $y(x) = e^x$

(artificial)

We have used $N = 10$, $\varepsilon = 10^{-14}$ and recorded below the maximum errors in the successive iterates.

Problem	error in iterate number						
	0	1	2	3	4	5	6
1	$0.5 \cdot 10^{-2}$	$0.5 \cdot 10^{-4}$	$0.5 \cdot 10^{-6}$	$0.5 \cdot 10^{-8}$	$0.5 \cdot 10^{-10}$	$0.5 \cdot 10^{-12}$	$0.5 \cdot 10^{-14}$
2	$1.2 \cdot 10^{-2}$	$1.9 \cdot 10^{-4}$	$3.0 \cdot 10^{-6}$	$4.3 \cdot 10^{-8}$	$6.4 \cdot 10^{-10}$	$9.3 \cdot 10^{-11}$	$8.7 \cdot 10^{-11}$

Note 1 With $N = 10$ one cannot expect to get the error less than $O(h^{11})$, which is obtained after five iterative improvements. For problem 1, however the sixth iterate improves the accuracy as much as the previous iterates. This astonishing result is due to the fact that in the error expansion

$$\phi_{h,v}(\Delta_h y) = \Delta_h^0 \{F(y) + \sum_{j=(v+1)\cdot 2}^{M} h^j \psi_j(y)\} + O(h^{M+1})$$

we have $\psi_j(y) = 0$ for the exact solution y of the integral equation. This happens because the quadrature rules that are used for $v = 1, 2, \ldots$ are exact for polynomials of degree three or less and for the exact solution the integrand is t^3. cf. note 1 after the numerical results of section 4.3.1.

5. Concluding remarks

In the final stage of the work with this report we got a preliminary version of a paper, "Iterated defect corrections based on estimates of the local discretization error," by R. Frank, J. Hertling and C. W. Uberhuber, report number 18/76 from the Institut fur Numerische Mathematik, Technical University of Wienna, Wienna, Austria. There the authors consider an algorithm very similar to our algorithm for iterative improvement. However their perturbation operators (or approximations to the local discretization error) are computed differently. For the two point boundary value problem

$$y'' = f(t, y) \qquad y(0) = A \qquad y(1) = B$$

with the discretization

$$\phi_h(\xi) = \begin{pmatrix} \xi_0 - A \\ \dfrac{\xi_{i+1} - 2\xi_i + \xi_{i-1}}{h^2} - f(t_i, \xi_i) = 0 \qquad i = 1, 2, \ldots, N-1 \\ \xi_N - B \end{pmatrix}$$

they define locally smooth functions $P_k^0(t, \eta^0)$, $k = 1, 2, \ldots, N-1$ that interpolate the solution η^0 of $\phi_h(\eta^0) = 0$ at some points surrounding $x = x_k$. Then they compute the local discretization error as

$$\ell_k(\eta^0) = (P_k^0(t_{k-1}, \eta^0) - 2P_k^0(t_k, \eta^0) + P_k^0(t_{k+1}, \eta^0))/h^2 - (P_k^0)''(t_k, \eta^0)$$

$$k = 1, 2, \ldots, N-1$$

The succeeding approximations to the local discretization error (corresponding to our approximations $\phi_{h,v}(\eta^{v-1})$, $v = 2, 3, \ldots$)are computed by the same technique, but now higher and higher order interpolation is used to define the functions $P_k^v(t, \eta)$. The polynomials $P_k^v(t, \eta^v)$ do not need to be computed explicitly but the second derivative can be computed by forming

linear combinations of the components of η^{ν}, as we do when we approximate linear functionals.

In our notation (cf. section 4.2) they define families of perturbation operators $\phi_{h,\nu}$, $\nu = 1, 2, \ldots$ according to

$$\phi_{h,\nu}(\xi) = \begin{pmatrix} \xi_0 - A \\ \dfrac{\xi_{k-1} - 2\xi_k + \xi_{k+1}}{h^2} - D_{t_k}^{2,2+2\nu}(\xi) \qquad k = 1, 2, \ldots, N-1 \\ \xi_{N-1} - B \end{pmatrix}$$

Note that for any $z \in E$

$$\phi_h(\Delta_h z) = \Delta_h^0\{F(z) + \sum_{j=1}^{M} f_j(z)\, h^{2j}\} + O(h^{M+1})$$

where

$$f_j(z) = \begin{pmatrix} 0 \\ \dfrac{2}{(2j+2)!}\, z^{(2j+2)} \\ 0 \end{pmatrix}$$

and

$$\phi_{h,\nu}(\Delta_h z) = \Delta_h^0\{\sum_{j=1}^{\nu} f_j(z)\, h^{2j} + \sum_{j=\nu+1}^{M} f_{\nu j}(z)\, h^{2j}\} + O(h^{M+1})$$

(if $D_{t_k}^{2,2+2\nu}$ uses points symmetrically distributed around t_k) with proper definitions of Δ_h and Δ_h^0. The improved solutions are computed according to

$$\phi_h(\eta^0) = 0$$

$$\phi_h(\eta^i) - \phi_{h,i}(\eta^{i-1}) = 0 \qquad i = 1, 2, \ldots$$

Our theory does not cover this type of perturbations, but theorems similar to ours could be proved with our technique for this algorithm.

References

[1] Ballester, C., Pereyra, V. (1967) "On the Construction of Discrete Approximations to Linear Differential Expressions," Math. Comp. 21, 297-302.

[2] Bellman, R. E., Kalaba, R. E. (1965) "Quasilinearization and Nonlinear Boundary Value Problems," American Elsevier Publishing Company, New York.

[3] Bjorck, A., Dahlquist, G. (1973) "Numerical Methods," Prentice-Hall, Englewood Cliffs, New Jersey.

[4] Bjorck, A., Pereyra, V. (1970) "Solution of Vandermonde Systems of Equations," Math. Comp. 24, 893-903.

[5] Ciarlet, P. G., Schultz, M. H., Varga, R. S. (1967) "Numerical Methods of High-Order Accuracy for Nonlinear Boundary Value Problems I," Num. Math. 9, 394-430.

[6] Cole, J. D. (1951) "On a Quasilinear Parabolic Equation Occurring in Aerodynamics," Quarterly of Applied Mathematics, Vol. IX, 225-236.

[7] Concus, P. (1967) "Numerical Solution of the Minimal Surface Equation," Math. Comp. 21, 340-350.

[8] Dahlquist, G., Lindberg, B. (1973) "On Some Implicit One-Step Methods for Stiff Differential Equations," TRITA-NA-7302, Dept. of Information Processing, The Royal Institute of Technology, Stockholm, Sweden.

[9] Fox, L. (1947) "Some Improvements in the Use of Relaxation Methods for the Solution of Ordinary and Partial Differential Equations," Proc. Roy. Soc. London Ser. A 190, 31-59.

[10] Frank, R. (1975) "The Method of Iterated Defect-Correction and Its Application for Two-Point Boundary Value Problems I," Report 8/75, Institute fur Numerische Mathematik, Technische Hochschule, Vienna.

[11] Galimberti, G., Pereyra, V. (1970) "Numerical Differentiation and the Solution of Multidimensional Vandermonde Systems," Math. Comp. 24, 357-364.

[12] Galimberti, G., Pereyra, V. (1971) "Solving Confluent Vandermonde Systems of Hermite Type," Num. Math. 18, 44-60.

[13] Gourlay, A. R., Morris, J. L. (1968) "Deferred Approach to the Limit in Nonlinear Hyperbolic Systems," Comp. J. 11, 95-101.

[14] Hofman, P. (1967) "Asymptotic Expansions of the Discretization Error of Boundary Value Problems of the Laplace Equation in Rectangular Domains," Num. Math. 9, 302-322.

[15] Joyce, D. C. (1971) "Survey of Extrapolation Processes in Numerical Analysis," SIAM Review 13, 435-490.

[16] Keller, H. B. (1970) "A New Difference Scheme for Parabolic Problems, in Numerical Solution of Partial Differential Equations II," SYNSPADE 1970 (ed. Hubbard B.), p. 327-350.

[17] Kolar, W. (1972) "Uber Differenzenverfahren Von Monotoner Art fur Nichtlineare Parabolische Randwert Probleme, in Numerische Losung Nichtlinearer Partieller Differential und Integro - Differential Gleichungen," R. Ansorge, W. Tornig (eds.), Springer Verlag.

[18] Kronsjo, L., Dahlquist, G. (1972) "On the Design of Nested Iterations for Elliptic Difference Equations," BIT 12, 63-71.

[19] Lambert, J. D. (1973) "Computational Methods in ODE's," J. Wiley & Sons, London.

[20] Laurent, P. J. (1964) "Etudies des Procedes d'Extrapolation en Analyse Numerique," Thesis, Univ. of Grenoble, Grenoble, France.

[21] Lentini, M., Pereyra, V. (1974) "A Variable Order Finite Difference Method for Nonlinear Multipoint Boundary Value Problems," Math. Comp. 28, 981-1003.

[22] Lentini, M., Pereyra, V. (1975a) "Boundary Problem Solvers for First Order Systems Based on Deferred Corrections, in Numerical Solution of Boundary Value Problems for Ordinary Differential Equations," Academic Press, New York.

[23] Lentini, M., Pereyra, V. (1975b) "An Adaptive Finite Difference Solver for Nonlinear Two-Point Boundary Problems with Mild Boundary Layers," Rpoert STAN-CS-75-530, Computer Science Dept., Stanford University.

[24] Lister (1960) "The Numerical Solution of Hyperbolic Partial Differential Equations by the Method of Characteristics, in Mathematical Methods for Digital Computers," A. Ralston and H. S. Wilf (eds.), J. Wiley, New York.

[25] Moore, R. H. (1968) "Approximation to Nonlinear Operator Equations and Newton's Method," Num. Math. 12, 23-34.

[26] Netravali, A. N., de Figueiredo, R. J. P. (1974) "Spline Approximation to the Solution of the Linear Fredholm Equation of the Second Kind," SIAM J. Numer. Anal. 11, 538-549.

[27] Pereyra, V. (1967a) "Iterated Deferred Corrections for Nonlinear Operator Equations," Num. Math. 10, 316-323.

[28] Pereyra, V. (1967b) "Accelerating the Convergence of Discretization Algorithms," SIAM J. Numer. Anal. 4, 508-533.

[29] Pereyra, V. (1968) "Iterated Deferred Corrections for Nonlinear Boundary Value Problems," Num. Math. 11, 111-125.

[30] Pereyra, V. (1970) "Highly Accurate Numerical Solution of Casilinear Elliptic Boundary Value Problems in n Dimensions," Math. Comp. 24, 771-783.

[31] Pereyra, V. (1973) "High Order Finite Difference Solution of Differential Equations," Report STAN-CS-73-348, Computer Science Dept., Stanford University.

[32] Rall, L. B. (1969) "Computational Solution of Nonlinear Operator Equations," J. Wiley & Sons, New York.

[33] Skollermo, G. (1975a) "How the Boundary Conditions Affect the Stability and Accuracy of Some Implicit Methods for Hyperbolic Equations," Report No. 62, Dept. of Computer Science, Uppsala University, Uppsala, Sweden.

[34] Skollermo, G. (1975b) "Error Analysis for the Mixed Initial Boundary Value Problem for Hyperbolic Problems," Report No. 63, Dept. of Computer Science, Uppsala University, Uppsala, Sweden.

[35] Smith, R. R. (1970) "Extrapolation Applied to the Numerical Solution of Hyperbolic Partial Differential Equations," Doctoral Thesis, University of California, San Diego.

[36] Stetter, H. J. (1965) "Asymptotic Expansions for the Error of Discretization Algorithms for Nonlinear Functional Equations," Num. Math. 7, 18-31.

[37] Stetter, H. J. (1973) "Analysis of Discretization Methods for Ordinary Differential Equations," Springer Verlag, New York.

[38] Stetter, H. J. (1974) "Economical Global Error Estimation, in Stiff Differential Systems," R. A. Willoughby (ed.), Plenum Press, New York.

[39] Van der Sluis, A. (1972) "The Remainder Term in Quadrature Formulae," Num. Math. 19, 49-55.

[40] Volkov, E. A. (1957) "A Method for Improving the Accuracy of Grid Solutions of the Poisson Equation (in Russian)," Vycisl. Mat. 1, 62-80, translation in American Mathematical Society Translations, Ser. 2, Vol. 35 (1964).

[41] Werner, W. (1968) "Numerical Solution of Systems of Quasilinear Hyperbolic Differential Equations by Means of the Method of Neben-Characteristics in Combination with Extrapolation Methods," Num. Math. 11, 151-169.

BIBLIOGRAPHIC DATA SHEET	1. Report No. UIUCDCS-R-76-820	2.	3. Recipient's Accession No.
4. Title and Subtitle: Error Estimation and Iterative Improvement for the Numerical Solution of Operator Equations			5. Report Date: July 1976 6.
7. Author(s): Bengt Lindberg			8. Performing Organization Rept. No. UIUCDCS-R-76-820
9. Performing Organization Name and Address: Department of Computer Science University of Illinois at Urbana-Champaign Urbana, IL 61801			10. Project/Task/Work Unit No. 11. Contract/Grant No. AFOSR-75-2854
12. Sponsoring Organization Name and Address: Department of the Air Force Air Force Office of Scientific Research (AFSC) Bolling Air Force Base, DC 20332			13. Type of Report & Period Covered 14.

15. Supplementary Notes

16. Abstracts

A method for estimation of the global discretization error of solutions of operator equations is presented. Further an algorithm for iterative improvement of the approximate solution of such problems is given. The theoretical foundation for the algorithms are given as a number of theorems. Several classes of operator equations are examined and numerical results for both the error estimation algorithm and the algorithm for iterative improvement are given for some classes of ordinary and partial differential equations and integral equations.

17. Key Words and Document Analysis. 17a. Descriptors

operator equations, discretization error, iterative improvement

7b. Identifiers/Open-Ended Terms

7c. COSATI Field/Group

18. Availability Statement	19. Security Class (This Report)	21. No. of Pages
Unlimited	UNCLASSIFIED	94
	20. Security Class (This Page) UNCLASSIFIED	22. Price